First publ. in: Journal of New Music Research 37 (2008), 3, pp. 233–245

The Perception of Artistic Quality in Opera – Results from a Field Study

Sabine Boerner and Johanna Jobst

University of Konstanz, Germany

Abstract

The authors focus on spectators' judgements on the performance quality in opera. The results of a field study conducted in Dessau Opera House revealed the single components (e.g. orchestra) and the congruency components (e.g. congruency between the music and the staging dimension) which contribute to operagoers' overall quality judgements. Spectators' individual judgements were found to be highly homogeneous, with only minor differences between experts and non-experts. Implications for the production and the management of opera are discussed.

1. Introduction

Although operagoers regularly pronounce judgements on the artistic quality of an opera performance, the emergence of quality judgements in opera has rarely been analysed (Balme, 2003). Systematic knowledge of the emergence of quality judgements in opera would allow one to further elucidate the audience's reactions to a performance. It is hence essential for both artists and managers in opera.

From the artistic perspective, research on the emergence of quality judgements could help understand audience feedback. Belonging to the performing arts, opera is inherently geared to the audience. Those involved in the production of opera thus cannot go without feedback from their audience. However, feedback by the audience is usually limited to spontaneous reactions during and after the performance and to professional reviews. Further analysing the emergence of quality judgements in opera would thus provide members of the ensemble with systematic insights in what spectators perceive, think and feel during a performance and how they evaluate their observations. Members of the ensemble like orchestra musicians and soloists, for example, may be interested to know how their particular contribution to the performance is appreciated by the audience. Similarly, stage directors, dramaturges, and conductors may be interested in the relative weight which spectators give to the staging and to the musical aspects of a production, respectively.

From the management perspective, knowledge of the emergence of spectators' quality judgements is essential to intensify customer orientation. Following Kotler and Scheff (1997, p. 34), customer orientation "requires that the organization systematically study customers' needs and wants, perceptions and attitudes, preferences and satisfaction". Similarly, Rentschler et al. (2002, p. 122) argued, "the more the organization learns about and monitors the patrons' needs, preferences, attitudes and concerns, the more their satisfaction and commitment levels grow". Research on audience judgements in opera could thus contribute to answer the question *why* some productions or performances are appreciated by the audience while others are not. Moreover, investigating the role of spectators' personality (e.g. gender, age, expertise in opera) for their quality judgements, audience research in opera could support opera companies to identify different segments of customers and to decide who should be involved in the evaluation of performance quality.

The paper thus sheds light on the emergence of quality judgements in opera. Our first goal is to identify the

Correspondence: Prof. Dr. Sabine Boerner, Universität Konstanz, Fach 88, Universitätsstrasse 10, 78457 Konstanz, Germany.
E mail: Sabine.Boerner@uni konstanz.de

criteria that spectators in opera apply to generate their overall judgements on the performance quality. Opera is multi-dimensional in that the presentation of music, language/plot, and manifold visual stimuli is combined. Criteria for spectators' quality judgements in opera may thus be identified by looking at approaches in both music psychology and theatre studies. Research in music psychology is limited to a range of instruments to gauge the listeners' quality judgements on selected aspects of an opera performance that might be relevant for the overall quality judgement: individual players of instruments (e.g. Wapnik & Ekholm, 1997), individual singers (e.g. Kleber, 2004), and orchestras and choirs (e.g. Sagen, 1983). Similarly, theatre studies suggested scales for the evaluation of acting quality (Konijn, 1992) and criteria for the analysis of theatre performances to be used in students' education (Pavis, 1988). However, audience research in theatre so far did not provide a comprehensive approach to understand spectators' quality judgements. Against this background our first research question is: *Which components of an opera performance contribute to operagoers' overall quality judgements?*

The second aim of our study is to analyse the impact *situational and personal conditions* may have on the generation of individual judgements on performance quality in opera. On the one hand, spectators' judgements on the performance are based on subjective expectations and perceptions (Thompson & Williamon, 2003) and may therefore differ from person to person. In the arts management literature (e.g. Kotler & Scheff, 1997), several personal conditions are discussed that may influence individual judgements (e.g. age, gender, education, income, family status, lifestyle). On the other hand, attending a live performance inherently is a collective experience (Eversmann, 2004), offering the chance of homogenizing interpersonal influences on individual judgements. Applying these contradictory results to the context of opera, our second research question is: *To what extent do individual spectators agree in their judgements on the artistic quality of an opera performance?*

In the literature, expertise is assumed to influence the perception and evaluation of theatre and music performances (e.g. Holbrook, 1999; Boorsma & van Maanen, 2003). Accordingly, our third research question is: *Do ratings of quality in an opera performance differ as a function of spectators' expertise in opera?*

2. Components of operagoers' judgements on performance quality

Combining research on music psychology and theatre studies, Boerner (2004) suggested a componential framework of performance quality in opera (see Table 1). In this model, a music dimension (orchestra, choir, soloists) is distinguished from a staging dimension (acting quality, staging quality of a performance in opera). Both the music dimension and the staging dimension include a set of components that can be further specified. In the music dimension, the component "orchestra", for example, can be differentiated by instrument groups (strings, wind, percussion), for each group by individual instruments (e.g. violin, viola), and for each instrument by part (1st and 2nd violin). The choir as another component can be differentiated by voice (soprano, alto, tenor, and bass) and part (e.g. 1st and 2nd soprano). On the staging dimension, the acting quality can be divided into the performer's activity (language, expression, gestures, and movement) and his/her appearance (makeup and hairstyle). Staging quality includes costumes and stage set (spatial conception and stage space).

In addition, due to the multi-dimensional nature of the stage work (Eversmann, 2004) and the complexity of the differentiated dimensions, judging artistic quality in opera goes beyond the isolated components. Equally important for the judgement on performance quality is the perceived congruity between and within these components (Adorno, 1968; Eversmann, 2004). In Boerner's (2004) model, congruity is specified on three levels and referred to as "fit" (cp. North & Hargreaves, 1997).

Firstly, fit between the music and the staging dimension of a performance (*first-order fit* or "*fit1*") is important. For example, the tonal image produced in the music dimension should "match" the atmosphere conveyed in the staging dimension. The criteria for the first-order fit of an opera performance depend on the guiding idea which the specific artistic conception takes as the basis for a production. Secondly, fit must be achieved within each dimension (*second-order fit* or "*fit2*"). Within the music dimension, the individual components (orchestra, choir, solo pieces) should harmonize with respect to sound (e.g. intonation, articulation, dynamics). The same is true for the components of the staging dimension: the acting quality and the costumes, for example, should harmonize with respect to the intended atmosphere. Thirdly, fit applies to the given components of a dimension, for example, within the orchestra or within staging quality (*third-order fit* or "*fit3*"). In the orchestra the entrance of the individual musicians must be synchronized (Boerner & von Streit, 2006). Similarly, costumes, spatial concept, and stage space must be coordinated with the overall staging conception.

According to this framework, the construct "performance quality in opera" can be conceptualized as multi-dimensional, including the following aspects (see Table 1): orchestra (including "fit3"), choir_music (including "fit3"), choir_staging, soloists_music (including "fit3"), soloists_staging, scenery (including "fit3"), fit within the music dimension ("fit2_music"), fit within the staging dimension ("fit2_staging"), and fit between

Table 1. Componential framework of performance quality in opera.

First-order fit (fit 1) Fit between music dimension and staging dimension	
Music dimension	Staging dimension
Second-order fit (fit 2) Fit within the music dimension	Second-order fit (fit 2) Fit within the staging dimension
Orchestra quality (instruments, parts)	Quality of the choir (parts)	Quality of the soloists' parts	Acting quality (behaviour, appearance)	Scenery quality (costumes, spatial concept)
Third-order fit (fit 3)	Third-order fit (fit 3)	Third-order fit (fit 3)	Third-order fit (fit 3)	Third-order fit (fit 3)

staging dimension and music dimension ("fit1"). In order to check if operagoers *actually* use these aspects for their judgements on the performance quality in opera, a content-analysis of expert reviews of opera performances found in nationwide newspapers was conducted, showing support for the model (Boerner et al., 2008).

In our study, we used this componential framework to answer our first research question. We thus assume that the components specified in the model contribute to spectators' judgements on the performance quality in opera.

Hypothesis 1. *The perceived performance quality in opera is a function of spectators' perceptions of single components (quality of orchestra, choir, soloists, scenery) and perceptions of congruity components (fit between staging dimension and music dimension, fit within the staging dimension, fit within the music dimension).*

3. Intersubjective (dis-)agreement on performance quality

As Eversmann (2004) showed for theatre plays, the subjective experience of a performance includes the personal involvement, the perceptions, thoughts, and emotions that arise while watching the event. Thus, subjective judgements in opera are likely to be influenced by the spectator's individual traits. More precisely, extant literature in music psychology, theatre studies, empirical aesthetics, and perception in the performing arts identified the recipient's gender, education, prior experience, prior information, theatre- and music-related norms and values, state of mind, and personal taste to be relevant predictors of individual quality judgements (e.g. Kotler & Scheff, 1997; Boorsma & van Maanen, 2003; Thompson & Williamon, 2003). To the extent that these results are transferable to the assessment of performances in opera, arguments can be found both for a high and a low intersubjective agreement in quality judgements among operagoers.

On the one hand, the large number of interdependent aspects presented during an opera performance and their interplay (see Table 1) create a considerably complex experience (cp. Berlyne, 1971). Confronted with this complex stimulus, spectators are likely to experience sensual overload (Eversmann, 2004). In order to reduce the information flow, they could attend less strongly to several aspects, thus letting the opera 'wash over' them. Alternatively, they could reduce the complexity by focussing their attention on selected aspects (Broadbent, 1964) of the performance (e.g. the choir, the orchestra, the set) while neglecting others. Since this selection is likely to be guided by spectators' individual preferences and expectations (Kotler & Scheff, 1997; Thompson &

Williamon, 2003), differences between individual judgements are to be expected.

In addition, according to their individual expectations and preferences, individual spectators may tend to give different *weight* to the aspects they apply in their overall judgement. Whereas one spectator may give particular weight to the soloists' performance, his/her neighbour may base his/her overall judgement mainly on the performance of the orchestra. These differences in weighting the aspects of an opera performance are likely to result in differences in individual judgements on the performance. Furthermore, individual spectators may differ in their individual application of the evaluation criteria. For instance, visitor A's understanding of a virtuous soloist (e.g. a brilliant baritone voice) may differ from visitor B's understanding. While both are listening to, for example, Papageno, both applying the criterion of voice brilliance, they may come to different judgements about Papageno's performance quality.

In sum, individual differences in visitors' expectations and preferences are likely to result in a generally low interrater agreement in operagoers' judgements on the performance quality.

Hypothesis 2a. *Visitors' intersubjective agreement on the performance quality in opera will be low.*

On the other hand, research has revealed a tendency towards demographic homogeneity of consumers who regularly attend high culture art events (Johnson & Garbarino, 2001). In general, the culture-consuming public is more educated, has higher incomes, and has higher status jobs than the general public (Sargeant, 1997). Moreover, since opera belongs to the so-called highbrow art forms (Katz-Gerro, 2002), members of the audience in opera are likely to belong more or less to the same class. Aligning individual expectations and preferences, demographic homogeneity may thus contribute to homogenize individual quality judgements.

In addition, the attendance of a live opera performance inherently is a social experience. Therefore, homogenizing processes working in the audience have to be taken into account. Individual experiences may be influenced by group processes. By creating a common mood or climate (Joyce & Slocum, 1990) within the audience, interaction processes may result in repercussions for the individual spectators. As Eversmann (2004) claimed for theatre performances, spectators can be influenced by emotions of others in that their feelings may be intensified and reinforced by the reactions of their neighbours. Additionally, group conformity (Asch, 1956) has been found to influence aesthetic judgements in general (Crozier, 1996). Since individuals depend on the group for social approval and acceptance, they comply with the group because they anticipate being rewarded for doing so or being punished for not doing so. Group conformity may thus induce spectators to make dependent quality assessments, modifying their individual positions according to the majority opinion. Interactions among members of the auditorium, sometimes explicitly manifested in laughter, ''boos'' or ''bravos'', may thus have a harmonizing effect on quality evaluation, resulting in what Boorsma and van Maanen (2003, p. 329) called ''collective perceptions''. The resulting quality evaluations may thus be more homogenous than in the case of isolated assessments.

Hypothesis 2b. *Visitors' intersubjective agreement on the performance quality in opera will be high.*

4. Differences between experts and non-experts in their quality judgements

The assumption of a cultural hierarchy of quality evaluations in the arts (e.g. Bourdieu, 1993; for a review see Holbrook, 1999) suggests that experts in opera may differ from non-experts in their quality judgements. Because of their connoisseurship, experts establish and follow conventionally approved standards to their judgements. In other words, they know what is considered ''good'' according to the criteria sanctioned by the relevant cultural field. Therefore, experts are recognized and legitimated as arbiters of ''good taste'' in a particular cultural field (Holbrook, 1999). Accordingly, some argue that measuring the quality of an artistic performance can best be accomplished by appropriately experienced evaluators who are able to be objective and unbiased (cp. Thompson & Williamon, 2003).

In contrast, non-experts or ordinary attendees apply the standards of popular appeal instead of sharing the professional standards of evaluation (Holbrook, 1999). Whereas professional standards are autonomous or internal to the field, those of popular appeal are related to market success and therefore heteronymous or external to the field. Because of their ''bad taste'', non-experts are less likely than experts to produce valid and reliable judgements on artistic quality. Accordingly, empirical evidence from various cultural fields indicates only weak positive associations between expert and non-expert judgements (cp. Holbrook, 2005).

However, empirical studies of assessment of music performance (Burnsed et al., 1985; De la Motte-Haber & Rötter, 2005), solo vocal performance (Wapnik & Ekholm, 1997; Kleber, 2004), and theatre critics (Boorsma & van Maanen, 2003) have produced contradictory results. While critics' ratings of theatre performances have been found to differ significantly from those of non-experts (Boorsma & van Maanen, 2003), studies of orchestral performance that have explicitly compared

evaluators with differing levels of specialist training (e.g. Winter, 1993; Thompson, 2006) found only few notable differences between experts and non-experts.

In the field of opera, differences between experts' and non-experts' judgements on performance quality may occur concerning several aspects. First, experts may differ from non-experts in the weight they give to the individual components in their overall quality judgements. The professional standards which experts rely on may include conventions regarding these emphases, such as giving generally high or low weight to the staging dimension as opposed to the music dimension. Non-experts, however, may rely on their personal impression of the performance. If, for example, they are particularly impressed by the staging, this may result in a high relevance of the staging dimension for their overall quality judgements. Moreover, non-experts may have difficulties in assessing one or the other aspect, resulting in lower weight given to this aspect in their overall judgements.

Hypothesis 3a. *Experts in opera differ from non-experts regarding the weight that is given to the determinants of their overall quality judgement.*

Second, experts are likely to produce more homogenous judgements than non-experts. According to Bourdieu (1993), one could argue that agreement within an expert spectator group will be higher than agreement within a non-expert spectator group. Due to their professional identity (Tajfel, 1981), experts in opera will try to base their judgements explicitly on established professional standards. Thus, whereas non-professionals mainly take into account their individual tastes, experts "write and judge in the context of a field of professional colleagues" (Boorsma & van Maanen, 2003, p. 327). This may lead to homogenize experts' judgements on an opera performance more than non-experts' judgements.

Hypothesis 3b. *Experts in opera pronounce more homogenous quality judgements than non-experts.*

Third, experts may be more rigorous in their judgements on an opera performance, giving lower quality ratings than non-experts do. Due to their professional background, experts in theatre are assumed to refer to a higher level of expectation and thus to make more deliberate, critical evaluations than non-experts do (Boorsma & van Maanen, 2003). As Thompson and Williamon (2003) found for music performance assessments, evaluators gave lower ratings to the instruments which they are familiar with, probably because of their more detailed knowledge of the relevant artistic and technical issues. Applying this reasoning to judgements on an opera performance, it is likely that experts generate more rigorous quality judgements than members of the ordinary audience.

Hypothesis 3c. *Experts generate more rigorous quality judgements than non-experts.*

5. Method

5.1 Measures

5.1.1 Performance quality in opera

To measure performance quality in opera, we used scales from the "questionnaire for the perception of performance quality in music theatre" that has been developed and validated by Boerner et al. (2008). Based on the componential framework on performance quality in opera (see Table 1), this questionnaire resulted from qualitative content-analyses of media reviews of opera-performances and quantitative pilot studies. Similar to research in music psychology (Sagen, 1983), five point Likert-Scales were used, with 1 indicating the lowest ("strongly disagree") and 5 indicating the greatest quality ("strongly agree"). In order to assure that the responses to the questionnaire reflect the results of the audience's experience in opera rather than being a result of the questions asked and the way they were asked, two additional response options were provided. To check if and how far the questionnaire corresponded to the internal marking schemes spectators really apply (Thompson & Williamon, 2003, p. 35) and to avoid arbitrary answers in case of respondents' uncertainty, the response scale was enlarged by the alternatives "I did not pay attention to" and "I am not able to judge".

The questionnaire included the following scales. Scales measuring the individual components of an opera performance: orchestra, choir_music, choir_staging, soloists_music, soloists_staging, and scenery. In order to avoid an average judgement on the soloists' quality, our instruction was to concentrate exclusively on one soloist;[1] scales capturing the congruity ("fit") within and between these components ("*fit1*", "*fit2_music*", "*fit2_staging*"; see Table 1) and a scale capturing the overall performance. Examples of the measurement of these constructs are given in Table 2.

5.1.2 Expertise in opera

Expertise is generally understood as competence in a particular field, including both specialized knowledge and experience (e.g. Ericsson et al., 1993). In our study,

[1]Since we analysed a life performance of "The Magic Flute", participants were asked to concentrate exclusively on "Papageno".

Table 2. The questionnaire for the perception of performance quality in music theatre constructs and exemplary items.

constructs (number of items)	exemplary items
overall artistic quality of the performance (3)	Altogether, I was absolutely convinced of the artistic quality of the performance.
fit1 (5)	Altogether in this performance the "music" (orchestra, choir, and soloists) matched the "scene" (stage settings, decoration, props, costumes, lighting) very well.
fit2_music (5)	With respect to the volume, orchestra, choir and soloists fitted very well.
fit2_staging (8)	Costumes and stage setting matched in such a way that a "harmonious" atmosphere was developed.
orchestra (10)	The orchestra elaborated differences in volume very well.
soloists_music (5)	The soloist's voice was little beautiful in terms of sound. R
soloists_staging (4)	The soloist embodied the figure's characteristics very convincingly.
choir_music (7)	The choir accentuated differences in volume very well.
choir_staging (3)	From its appearance, one could fully believe the choir's role.
scenery (3)	I was highly impressed by the stage setting.

Note: Scenery = general stage setting; fit2_music = interplay between orchestra, soloists and choir; fit2_staging = interplay between the staging aspects; fit1 = interplay between music dimension and staging dimension; R = polarity reversed; five point Likert scales were used, with 5 indicating the greatest quality.

we chose to capture the respondents' expertise accordingly, using self-report measures. To increase the grade of accuracy specialized knowledge in opera was not directly asked for, but operationalized by one item concerning the knowledge of the piece ("How well did you know the performed opera before?") and one item asking for specialized knowledge of music ("Do you play an instrument?"). For both items, we used five-point Likert scales ("strongly agree, agree, neither agree nor disagree, disagree, strongly disagree"). Experience with the opera was measured by two items ("How much experience do you have with the opera?": "complete layman, little knowledge, hobbyist, semi-professional, expert" and "How often do you visit the opera during a year on average?"). These four items were added to an index which divided the sample into two sub-samples, referred to in this paper as experts ($n = 57$) and non-experts ($n = 57$).

5.1.3 Demographic variables

Demographic variables (age, sex, and general education) that have proved to be relevant for evaluation in the performing arts (DiMaggio et al., 1978) were included.

5.2 Sample

The study was conducted during a 2007 live performance of "The Magic Flute" at Dessau Opera House (conducted by Golo Berg and directed by Johannes Felsenstein). After an announcement of the study had been made, the questionnaires were distributed randomly among audience members. 114 questionnaires were completed resulting in a response rate of approximately 35% (Jobst, 2007). In comparison to similar surveys (Reuband, 2005) and considering the time restriction when collecting data in theatres (Roose et al., 2003), this can be seen as an acceptable value. Since participants were predominantly female (58%) and 49 years on average, our sample is representative for a typical opera audience (Sargeant, 1997).

6. Results

6.1 Preliminary analyses

As was expected, Cronbach's alphas were all above 0.80, showing sufficient reliability for all variables (see Table 3). Based on these results, summated scales were built by adding the questionnaire items measuring the same variable (Vogt, 1993). For every scale, each respondent's answers were added up to arrive at a single number measuring the strength of his or her attitude. Subsequently, these summated scales were divided by the number of items, respectively, and used for the analyses in the remainder of the study. The intercorrelations of all variables under study are shown in Table 3. Similar to previous studies (Boerner et al., 2008), factor analyses suggested to summarize "soloists_music" and "soloists_staging" into one comprehensive factor.

Descriptive analyses of the response options "I did not pay attention to" and "I am not able to judge" revealed that less than 3% of the respondents chose one of these alternatives ("I did not pay attention to": 0.8%; "I am not able to judge": 2.0%). As there was no item for which more than 10% of the respondents chose these response options, we decided not to exclude any (Roth, 1994). Hence, the questionnaire seems to represent spectators' internal marking schemes pretty well.

Table 3. Intercorrelations between the constructs and reliabilities (Cronbach's alpha).

	1.	2.	3.	4.	5.	6.	7.	8.	9.	Mean	SD
1. overall artistic quality of the performance	(0.90)									3.87	1.02
2. fit 1	0.78**	(0.92)								4.03	0.86
3. fit2_music	0.58**	0.52**	(0.82)							4.30	0.69
4. fit2_staging	0.62**	0.80**	0.35**	(0.95)						3.86	0.95
5. orchestra	0.73**	0.50**	0.64**	0.32**	(.97)					4.42	0.76
6. soloists	0.76**	0.68**	0.61**	0.50**	0.78**	(0.90)				4.28	0.65
7. choir_music	0.65**	0.53**	0.53**	0.42**	0.76**	0.74**	(0.93)			4.43	0.66
8. choir_staging	0.41**	0.58**	0.25*	0.56**	0.25*	0.41**	0.42**	(0.94)		3.63	1.12
9. scenery	0.64**	0.71**	0.25**	0.83**	0.39**	0.49**	0.43**	0.44**	(0.88)	3.84	1.05

Note: $^{**}p \leq 0.01$, $^{*}p \leq 0.05$; Scenery = general stage setting; fit2_music = interplay between orchestra, soloists, and choir; fit2_staging = interplay between the staging aspects; fit1 = interplay between music dimension and staging dimension; Reliabilities (Cronbach's Alpha) shown on the diagonal.

6.2 Components of the judgement on an opera performance

In order to investigate which components contribute to visitors' quality judgements (Hypothesis 1), structural equation modelling (SEM) was employed (AMOS 7.0; Kline, 2005). Combining the properties of factor analysis, regression analysis, and path analysis, SEM enables the definition and estimation of complex model structures, thus offering the potential to account for multiple influences, which may simultaneously affect various outcome variables (e.g. the performance of the soloist simultaneously influencing the fit within the music dimension and the fit within the staging dimension). Furthermore, SEM allows for the inclusion of latent variables which do not suffer from systematic restrictions in measurement quality, because measurement error is explicitly separated from true variance in the estimation process (Kline, 2005). The spectators' overall judgements on the performance of "The Magic Flute" were predicted using the components of the construct "performance quality in opera" (see Table 1). The SEM analysis confirmed that a hierarchical model ($\chi^2 = 3206.21$; $p \leq 0.001$; CMIN/DF = 2.46; NFI = 0.629; RMSEA = 0.111)[2] fits better to the data than an alternative general factor model (χ^2=4885.15; $p \leq 0.001$; CMIN/DF = 4.16; NFI = 0.358; RMSEA = 0.17). Since we assumed that the overall performance may be determined by any of the components included in Table 1, we allowed *direct* paths from all components to the "perceived artistic quality".

The resulting model explained 80% of the variance in the spectators' overall judgements on the quality of the performance "The Magic Flute". In order to determine the contribution of the individual components to visitors' overall quality judgements, we calculated the total effects for these components (see Table 4). Interpreting these total effects, all assumed components contributed to visitors' overall quality judgements, except "choir_music". Hypothesis 1 was thus confirmed for all components except the musical performance of the choir.

6.3 Agreement within the audience on the artistic quality of an opera performance

To investigate if and to what degree individual evaluations of the performance "The Magic Flute" differ (Hypotheses 2a and 2b), we calculated the interrater agreement within the audience. In the literature, different indices to determine the interrater agreement are discussed (e.g. Chen & Krauss, 1993). Based on current recommendations (e.g. Chen & Krauss, 1993; von Eye & Mun, 2005), we decided to calculate the intraclass correlation coefficient ICC_{unjust} (Commenges & Jacqmin, 1994; Wirtz & Caspar, 2002).[3] The attendees of "The Magic Flute" highly agreed on all constructs of the artistic quality (see Table 5), confirming Hypothesis 2b.

6.4 Differences between experts and non-experts

Since the expert subsample and the non-expert subsample of our study were too small to conduct separate structural equation modelling (Kline, 2005), we calculated path analyses to investigate if and to what degree experts and non-experts differ regarding the weight they give to the determinants of the quality judgements (Hypothesis 3a). The resulting models explained an

[2] χ^2 (CMIN) = Chi Square Value, p = probability level, df = degrees of freedom, NFI = Normed Fit Index, RMSEA = Root Mean Square Error of Approximation (Schermelleh Engel et al., 2003).

[3] A sufficiently high interrater agreement is assumed if $ICC_{unjust} = .70$ is met (Wirtz & Caspar, 2002).

Table 4. Standardized total effects of the components of the perceived performance quality in opera (results from structural equation modelling).

	Independent							
Dependent	scenery	choir_staging	choir_music	soloists	orchestra	fit2_staging	fit2_music	fit1
fit2_staging	0.70	0.27	0.00	0.27	0.00	0.00	0.00	0.00
fit2_music	0.00	0.00	0.09	0.47	0.37	0.00	0.00	0.00
fit1	0.59	0.23	0.02	0.33	0.08	0.84	0.22	0.00
overall artistic quality of the performance	0.30	0.11	0.02	0.35	0.47	0.31	0.10	0.64

Note: Fit2_staging = interplay between the staging aspects; fit2_music = interplay between orchestra, soloists, and choir; fit1 = inter play between musical dimension and staging dimension.

Table 5. Intraclass correlation coefficient ICC_{unjust}.

Construct	ICC_{unjust}	sample size
overall artistic quality of the performance	0.886***	113
fit1	0.916***	100
fit2_music	0.822***	109
fit2_staging	0.946***	108
orchestra	0.970***	86
soloists	0.884***	102
choir_music	0.922***	95
choir_staging	0.940***	106
scenery	0.856***	112

Note: ***$p \leq 0.001$; scenery = general stage setting; fit2_ music = interplay between orchestra, soloists, and choir; fit2_staging = interplay between the staging aspects; fit1 = interplay between musical dimension and staging dimension; Varying sample size due to missing values.

equal amount of variance in both the experts' ($\beta = 0.81$) and the non-experts' ($\beta = 0.79$) overall judgements on the quality of the performance.

Considering the total effects in these models, only minor differences between experts and non-experts occurred (see Table 6). In both the expert and the non-expert sample, the congruity between the music dimension and the staging dimension ("fit1") was given the highest weight for the overall performance (total effect for experts = 0.42; total effect for non-experts = 0.72; see Table 6), followed by the scenery ($\beta = 0.52$) and the orchestra ($\beta = 0.46$) in the non-expert sample, whereas the experts considered orchestra ($\beta = 0.34$) and the congruity within the staging dimension ("fit2_staging", $\beta = 0.30$) on the second and third position, respectively. Thus, differences between experts and non-experts concerned mainly the position of "fit2_staging", which non-experts did not consider at all in their overall judgement on the performance quality while experts gave considerable weight to this aspect ($\beta = 0.30$).

In addition, experts' and non-experts' judgements differed with regard to the importance that "fit2_music" and "fit2_staging" had for the perceived congruity between these variables ("fit1"): in non-experts' judgements, "fit2_staging" ($\beta = 0.80$) was much more important for "fit1" than "fit2_music" ($\beta = 0.11$). This contrast was much smaller in experts' judgements ($\beta = 0.64$ for "fit2_staging" and $\beta = 0.39$ for "fit2_music"). Altogether, the experts in our sample seemed to consider the aspects for their judgements on the performance quality in a more balanced way than the non-experts. Hypothesis 3a was thus partially confirmed.

In order to test Hypothesis 3b, we calculated the interrater agreement separately among experts and among non-experts. Both experts and non-experts highly agreed on all constructs of the artistic quality (see Table 7). With the exception of "scenery", the values in the expert-sample group were slightly, but not significantly above the values in the non-expert sample. Our study thus did not reveal any significant difference between experts and non-experts concerning the agreement of their subjective judgements on opera, rejecting Hypothesis 3b.

T-tests calculated to investigate if experts pronounce more *rigorous* quality ratings than non-experts (Hypothesis 3c) revealed significant differences only for the components "orchestra" and "choir_music" (see Table 8). Hence, only for these components experts pronounced more rigorous judgements than non-experts. All other aspects of the performance "The Magic Flute" were equally assessed by the experts and the non-experts in our sample. Hypothesis 3c was thus only partially confirmed.

7. Discussion

7.1 Summary

This study analysed subjective judgements on performance quality in opera, addressing the following research

Table 6. Standardized total effects of the hierarchical path models for experts and non experts.

	Independent					
	fit2_staging		fit2_music		fit1	
Dependent	non experts	experts	non experts	experts	non experts	experts
fit1	.80	.64	.11	.39	.00	.00
overall artistic quality of the performance	.00	.30	.19	.25	.72	.42

Note: Fit2_music = interplay between orchestra, soloists, and choir; fit2_staging = interplay between the staging aspects, fit1 = inter play between music dimension and staging dimension.

Table 7. Interrater agreement among experts and non experts.

	ICC_{unjust}		sample size	
Scale	non experts	experts	non experts	experts
overall artistic quality of the performance	0.873***	0.893***	57	56
fit1	0.902***	0.933***	50	50
fit2_music	0.814***	0.827***	53	56
fit2_staging	0.940***	0.952***	52	56
orchestra	0.960***	0.968***	37	49
soloists	0.764***	0.919***	48	54
choir_music	0.858***	0.931***	41	54
choir_staging	0.939***	0.940***	50	56
scenery	0.880***	0.823***	56	56

Note: ***$p \leq 0.001$; fit1 = interplay between musical dimension and staging dimension; fit2_music = interplay between orchestra, soloists, and choir; fit2_staging = interplay between the staging aspects; Varying sample size due to missing values.

questions. (1) *Which components of an opera performance contribute to operagoers' overall quality judgements?* (2) *To what extent do individual spectators agree in their judgements on the artistic quality of an opera performance?* (3) *Do ratings of quality in an opera performance differ as a function of spectators' expertise in opera?*

Our hypotheses were tested in a survey of $n = 114$ attendees of a performance of "The Magic Flute" at Dessau Opera House. The results of our field study are as follows.

(1) In their overall judgement on the quality of an opera performance, spectators include both individual components (orchestra, soloists, choir, and staging) and fit-components ("fit1", "fit2_staging", and "fit2_music"), confirming our first hypothesis (with the exception of the component "choir_music"). In total, the structural equation model we calculated explained 80% of the variance in the spectators' overall judgements on the quality of the performance "The Magic Flute". This result suggested that we did not cover *all* the components contributing to the spectators' overall judgement on performance quality in opera, but a considerable part of them and much more variance in the quality judgements than is usually explained in audience research (Eversmann, 2004). Nevertheless, it is up to further research to discover additional predictors.

(2) Further, the results of our study confirmed Hypothesis 2b, suggesting that visitors' intersubjective agreement on the performance quality in opera was high. The high interrater agreement we found allows for the interpretation that differences to be expected from individual selecting, weighting and assessing the criteria used in judging the artistic quality of an opera performance have been cushioned by demographic homogeneity and homogenizing processes in the audience. As a consequence, the alternative Hypothesis 2a was rejected. In addition, in the case of "The Magic Flute" in Dessau, other homogenizing tendencies must be taken into account. Dessau Opera House is famous

Table 8. *t* Tests comparing experts and non experts.

Scale	*t* value	mean		standard deviation		sample size	
		non experts	experts	non experts	experts	non experts	experts
overall artistic quality of the performance	1.400	4.00	3.74	0.886	1.089	57	56
fit1	0.577	3.98	4.08	0.883	0.850	50	50
fit2_music	1.285	4.39	4.22	0.673	0.715	53	56
fit2_staging	0.683	3.79	3.92	0.960	0.967	52	56
orchestra	3.365**	4.71	4.23	0.460	0.842	37	49
soloists	1.198	4.38	4.23	0.438	0.758	48	54
choir_music	2.262*	4.59	4.33	0.379	0.751	41	54
choir_staging	1.208	3.51	3.77	1.082	1.137	50	56
scenery	0.181	3.82	3.85	1.137	0.947	56	56

Note: $^{**}p \leq 0.01$, $^{*}p \leq 0.05$; fit1 = interplay between musical dimension and staging dimension; fit2_music = interplay between orchestra, soloists, and choir; fit2_staging = interplay between the staging aspects; Varying sample size due to missing values.

for its director Johannes Felsenstein who continues the production style that was introduced by his father Walter Felsenstein at Berlin "Komische Oper" and called "music theatre" as opposed to "opera". Therefore, Dessau attracts operagoers who are specialized in this production style, which may have led to a self-selection bias (Berk, 1983) in the audience, resulting in homogenous expectation levels of all attendees. Furthermore, the production "The Magic Flute" in Dessau had received mainly favourable reviews in the media, which may also have contributed to homogenize expectation levels (Kotler & Scheff, 1997) and caused self-fulfilling prophecies (Merton, 1968) in the evaluation, which tend to assimilate individual quality judgements.

(3) Concerning differences between experts in opera and non-experts, we received mixed results. Hypotheses 3a, stating that experts would differ from non-experts in the weight given to the individual components, was partially confirmed. In the composition of their judgements, experienced operagoers seemed to pronounce more balanced judgements than inexperienced operagoers. In particular, non-experts put considerably lesser weight to the music dimension as opposed to the staging dimension than experts did. In the eyes of the non-experts, the music dimension of an opera performance may have been perceived as "given", as more or less "constant" because it is determined by the work. Non-experts probably considered the music dimension as such to be closer to the original work, while perceiving the staging dimension as an interpretation that may go beyond the original. Hence, in their judgements on the overall quality and on the fit between music dimension and staging dimension ("fit1"), non-experts focused primarily on the question if the staging dimension fits the music than the other way round. Instead, experts, relying on standards of the field, were used to consider all components in a more balanced way in their quality judgements.

Contrary to our expectation (Hypotheses 3b), experts' judgements on an opera performance were not more homogenous than non-experts' judgements. Obviously, in both the expert and the non-expert sample, homogenizing tendencies have been working. It can only be speculated if these homogenizing processes may have been driven by different effects. Whereas commitment to professional standards (Holbrook, 1999, 2005) is likely to homogenize the level of expectation and the selection of quality criteria in the expert sample, interaction processes during the performance (Joyce & Slocum, 1990; Eversmann, 2004) may have tempered differences in individual judgements in the non-expert sample.

The idea that experts will give lower ratings for an opera performance than non-experts (Boorsma & van Maanen, 2003) was only confirmed for judgements on the music dimension, namely the orchestra and the choir. Thus, Hypotheses 3c was only partially confirmed.

In sum, we found only minor differences between experienced and inexperienced operagoers in their judgements on the quality of the performance of "The Magic Flute". The present study thus only partially confirmed the assumption of a cultural hierarchy affecting quality evaluations for the context of opera. Instead, our findings were consistent with what Holbrook (2005, p. 77) called the "dignity-of-the-common-person-hypothesis": "People share the norms

for what is considered 'good' by those with expertise in a particular cultural field and thereby display aspects of 'good taste"'.

In the context of opera, one argument for this hypothesis may be that operagoers generally tend to have at least some experience with the opera (Behr, 1983). Thus, it is possible that all the respondents of our study had at least requisite minimal familiarity with the opera, resulting in a predominance of expert-like judgements of performance quality. Thus, all operagoers, to a greater or lesser extent experienced, may have shared the relevant standards of evaluation in the field to a certain degree. This argument leads over to limitations of our study and implications for further research.

7.2 Limitations and implications for further research

First, expertise in opera was measured by a self-rated scale, which could have resulted in some bias. Actually, descriptive analysis revealed that the majority (72%) of the non-experts in our sample reported to have "some experience" with opera or being "hobbyists" (7%). Further research may show whether the differences we expected between experienced and inexperienced operagoers can be confirmed if the sample include true laymen. Such research should include spectators who do not represent the "typical" audience in opera, i.e. the young and the less well educated or typical non-attendees.

This argument points at a second limitation of our study. Since our results were based on a single performance of "The Magic Flute" at Dessau Opera House, the degree to which these exploratory results can be generalized to assessment in opera is limited. While the selected work is "typical" in that it is the opera by far most often performed in Germany (Deutscher Buehnenverein, 2006), the Dessau production, although conventional, may probably not be regarded as "typical" (see above). The results regarding differences between experts and non-experts are persuasive in that they support prior findings (Boerner & Renz, 2008). Still, additional field studies are required in order to obtain more data that allow for sound conclusions and recommendations. Such studies would have to control for the genre of the presented work (e.g. "opera seria" or "opera buffa"), the type of performance (e.g. matinee, première, regular performance), the style of production (conventional versus modern or avant-garde), and the reputation of the opera company under study. In addition, our sample was self-selected in that only volunteers completed the questionnaire, which may have resulted in nonresponse problems (Roose et al., 2003). However, given time restrictions for data collection in theatres (Roose et al., 2003) and a restricted willingness to participate in a study late at night, this limitation is hard to avoid.

7.3 Implications for the production of opera

The results of our study underline that opera is a collective form of the performing arts, regarding both the production and the consumption. First, the congruence between music and staging ("fit1") was considered by far the most important component for the audience's evaluation of performance quality (see Table 4). This result may remind those involved in the production of opera, e.g. performers, directors, conductors, musicians, and dramaturges, of the extremely high task interdependence in opera and the high relevance of positive synergy for the resulting performance quality. Given the distinct heterogeneity in an opera ensemble concerning e.g. tenure, age, artistic and technical expertise (Boerner & von Streit, 2006) and the high level of individuality artists usually claim (Gaztambide-Fernandez, 2008), our study underlines the necessity of finding a balance between individualism and some kind of collectivism in the production of opera.

Second, spectators' judgements on the performance quality turned out to be highly unanimous. One explanation for this result could be that visitors' expectations regarding "The Magic Flute" at Dessau Opera House had been quite homogenous from the outset. An alternative or complementary explanation, however, would be that visitors' perceptions have been harmonized during the performance. If this is true, the notion of opera as a collective event (cp. Eversmann, 2004) receives confirmation from our data. Individuals' judgements on an opera performance are highly susceptible to influences from the audience. Producers of opera should thus not only consider the interaction between artists and spectators, but also interaction processes among the audience. If our result will be replicated in further research, the conclusion would be that either everybody or nobody in the audience will be enthusiastic about a performance in opera.

7.4 Implications for the management of opera

To date, performance measurement in professional opera companies has been limited to quantitative indicators. Commonly used criteria are, for example, audience attendance and subscriber levels, number of (debut) performances, number of new productions, and earned income (Voss & Voss, 2000; Brooks & Kushner, 2002). However, these indicators only allow for indirect conclusions regarding the artistic quality as perceived by the audience. Although a qualitative assessment of artistic performance has been claimed (e.g. Radbourne, 1998), neither management-focused nor culture-political research have provided a suggestion yet.

Our research hence presents an instrument that could be applied in opera companies in order to further explore their artistic quality as perceived by the audience.

Moreover, the results from our study contribute to intensify the companies' customer orientation, providing the following insights: first, our study is the first to allow for statements on the relative importance of different components of an opera performance for the spectators' evaluations. In order to meet customers' expectations, no component of the quality framework (see Table 1) should be neglected. The most important aspects for customers' judgements on performance quality are the congruence between music and staging (''fit1''), the orchestra, the soloists, and the staging.

Second, members of the audience highly agree in their evaluations of ''The Magic Flute''. Given the broad range of age (ranging from 18 to 75) and the heterogeneity in gender and education in our sample, this result is surprising. In particular, this finding suggests that the audience cannot be decomposed in customer segments (Kotler & Scheff, 1997). However, as mentioned above, the particularities of the well-known opera ''The Magic Flute'' may have contributed to homogenize customers' expectations in advance, thereby oppressing any difference between audience subgroups like ''the younger'' and ''the elder'' or men and women.

Third, experts' judgements in our sample differed from non-experts' judgements only slightly. Thus, in order to determine audiences' evaluations of performance quality in opera, opera management could involve experts and non-experts interchangeably.

References

Adorno, T. (1968). Reflexionen uber Musikkritik. In H. Kaufmann (Ed.), *Symposium für Musikkritik* (pp. 7–21). Graz: Institut fur Wertungsforschung,

Asch, S.E. (1956). Studies of independence and conformity: a minority of one against a unanimous majority. *Psychological Monographs*, *70*(9), 221–240.

Balme, C. (2003). *Einführung in die Theaterwissenschaft*. Berlin: Schmidt.

Behr, M. (1983). *Musiktheater – Faszination, Wirkung, Funktion*. Wilhelmshaven: Heinrichshofen.

Berk, R.A. (1983). An introduction to sample selection bias in sociological data. *American Sociological Review*, *48*(3), 386–398.

Berlyne, D. (1971). *Aesthetics and psychobiology*. New York: Appleton-Century-Crofts.

Boerner, S. (2004). Artistic quality in an opera company – towards the development of a concept. *Nonprofit Management and Leadership*, *4*(4), 425–436.

Boerner, S. & von Streit, C. (2006). Creating cooperative climate in an orchestra: the role of the musicians flow and the conductors leadership style. *Musicae Scientiae*, *X*(2), 243–261.

Boerner, S., Neuhoff, H., Renz, S. & Moser, V. (2008). Evaluation in music theater: empirical results on content and structure of the audience's quality judgment. *Empirical Studies of the Arts*, *26*(1), 15–35.

Boerner, S. & Renz, S. (2008). Performance measurement in opera companies – comparing subjective quality judgments of experts and non-experts. *International Journal of Arts Management*, *10*(3), 21–37.

Boorsma, M. & van Maanen, H. (2003). View and review in the Netherlands: the role of theatre critics in the construction of audience experience. *International Journal of Cultural Policy*, *9*(3), 319–335.

Bourdieu, P. (1993). *The field of cultural production*. New York: Columbia University Press.

Broadbent, D.E. (1964). *Perception and communication*. Oxford: Pergamon.

Brooks, A.C. & Kushner, R.J. (2002). What makes and arts Capital? Quantifying a city's cultural environment. *International Journal of Arts Management*, *5*(1), 12–23.

Burnsed, V., Hinkle, D. & King, S. (1985). Performance evaluations reliability at selected concert festivals. *Journal of Band Research*, *21*(1), 22–29.

Chen, P.Y. & Krauss, A D. (1993). Interrater agreement. In M.S. Lewis-Beck, A. Bryman & T.F. Liao (Eds.), *The Sage encyclopedia of social science research methods, volume 2* (pp. 511–513). Thousand Oaks: Sage Publications.

Commenges, D. & Jacqmin, H. (1994). The intraclass correlation coefficient: distribution-free definition and test. *Biometrics*, *50*(2), 517–526.

Crozier, R. (1996). Music and social influence. In A.C. North & D.J. Hargreaves (Eds.), *The social psychology of music* (pp. 67–83). Oxford: Oxford University Press.

De la Motte-Haber, H. & Rotter, G. (2005). Formwahrnehmung. In H. De la Motte-Haber & G. Rotter (Eds.), *Musikpsychologie* (pp. 263–267). Laaber: Laaber-Verlag.

Deutscher Buehnenverein (2006). *Wer spielte was? Werkstatistik 2004/05 des Deutschen Buehnenvereins*. Darmstadt: Mykenae Verlag.

DiMaggio, P., Useem, M. & Brown, P. (1978). *Audience studies of the performing arts and museums: a critical review*. Washington, DC: National Endowment for the Arts.

Ericsson, K., Krampe, R. & Tesch-Romer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, *100*(3), 363–406.

Eversmann, P.G.F. (2004). The experience of the theatrical event. In V.A. Cremona, P. Eversmann, H. van Maanen, W. Sauter & J. Tulloch (Eds.), *Theatrical events. Borders – dynamics – frames* (pp. 139–174). Amsterdam: Rodopi.

Gaztambide-Fernandez, R.A. (2008). The artist in society: understandings, expectations, and curriculum implications. *Curriculum Inquiry*, *38*(3), 233–265.

Holbrook, M.B. (1999). Popular appeal versus expert judgment of motion pictures. *Journal of Consumer Research*, *26*, 144–155.

Holbrook, M.B. (2005). The role of ordinary evaluations in the market for popular culture: do consumers have 'good taste'? *Marketing Letters*, *16*(2), 75–86.

Jobst, J. (2007). Evaluation in offentlichen Dienstleistungsorganisationen. Eine empirische Untersuchung zur Entstehung des Publikumsurteils im Musiktheater. Unpublished thesis, University of Konstanz, Germany.

Johnson, M.S. & Garbarino, E. (2001). Customers of performing arts organizations: are subscribers different from nonsubscribers? *International Journal of Nonprofit and Voluntary Sector Marketing*, *6*(1), 61–77.

Joyce, W.F. & Slocum, J.W. (1990). Strategic context and organizational climate. In B. Schneider (Ed.), *Organizational climate and culture* (pp. 130–150). San Francisco: Jossey-Bass.

Katz-Gerro, T. (2002). Highbrow cultural consumption and class distinction in Italy, Israel, West Germany, Sweden, and the United States. *Social Forces*, *81*(1), 207–229.

Kleber, B. (2004). Evaluation von klassischem westlichem Gesang. Unpublished master thesis, University of Konstanz, Germany.

Kline, R.B. (2005). *Principles and practice of structural equation modelling*. New York: Guilford Press.

Konijn, E.A. (1992). Waiting for the audience. In H. Schoenmakers (Ed.), *Performance theory – reception and audience research (advances in reception and audience research 3)* (pp. 157–182). Amsterdam: Tijdschrift voor Theaterwekenschap.

Kotler, R. & Scheff, J. (1997). *Standing room only*. *Strategies for marketing the performing arts*. Boston, MA: Harvard Business School Press.

Merton, R.K. (1968). *Social theory and social structure*. New York: Free Press.

North, A.C. & Hargreaves, D.J. (1997). Music and consumer behaviour. In D.J. Hargreaves (Ed.), *The social psychology of music* (pp. 268–289). New York and Oxford: Berghahn Books.

Pavis, P. (1988). *Semiotik der Theaterrezeption*. Tubingen: Narr.

Radbourne, J. (1998). Benchmarking performing arts centers in Australia. WCVM Journal and Resource Date Center.

Rentschler, R., Radbourne, J., Carr, R. & Rickard, J. (2002). Relationship Marketing, Audience Retention and Performing Arts Organisation Viability. *International Journal of Nonprofit and Voluntary Sector Marketing*, *7*(2), 118–130.

Reuband, K. (2005). Sterben die Opernbesucher aus? Eine Untersuchung zur sozialen Zusammensetzung des Opernpublikums im Zeitvergleich. In A. Klein & T. Knubben (Eds.), *Deutsches Jahrbuch für Kulturmanagement* (pp. 123–138). Baden-Baden: Nomos.

Roose, H., Waege, H. & Agneessens, F. (2003). Respondent related correlates of response behaviour in audience research. *Quality & Quantity*, *37*, 411–434.

Roth, P. (1994). Missing data: a conceptual review for applied psychologists. *Personnel Psychology*, *47*, S. 537–560.

Sagen, D.P. (1983). The development and validation of a university band performance rating scale. *Journal of Band Research*, *18*, 1–11.

Sargeant, A. (1997). Marketing the arts – classification of U.K. theater audiences. *Journal of Nonprofit & Public Sector Marketing*, *5*(1), 45–62.

Schermelleh-Engel, K., Moosbrugger, H. & Muller, H. (2003). Evaluating the fit of structural equation models: tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, *8*(2), 23–74.

Tajfel, H. (1981). *Human groups and social categories: Studies in social psychology*. Cambridge: Cambridge University Press.

Thompson, S. (2006). Audience responses to a live orchestral concert. *Musicae Scientiae*, *X*(2), 215–244.

Thompson, S. & Williamon, A. (2003). Evaluating evaluation: musical performance assessment as a research tool. *Music Perception*, *21*(1), 21–41.

Vogt, W.P. (1993). *Dictionary of statistics and methodology*. Newbury Park: Sage.

Von Eye, A. & Mun, E.Y. (2005). *Analyzing rater agreement – manifest variable methods*. Mahwah, NJ: Lawrence Erlbaum Associates.

Voss, Z.G. & Voss, G.B. (2000). Exploring the impact of organizational values and strategic orientation on performance in not-for-profit professional theatre. *International Journal of Arts Management*, *3*(1), 62–76.

Wapnik, J. & Ekholm, E. (1997). Expert consensus in solo voice performance evaluation. *Journal of Voice*, *11*, 429–436.

Winter, N. (1993). Music performance assessment: a study of the effects of training and experience on the criteria used by music examiners. *International Journal of Music Education*, *22*, 34–39.

Wirtz, M.A. & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität. Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen*. Bern: Hogrefe.