Görlitz, Katja; Sels, Tim

The Effect of Age Diversity in Groups on Peer Evaluations and Individual Performance

Kyklos

Provided in Cooperation with:
John Wiley & Sons

Suggested Citation: Görlitz, Katja; Sels, Tim (2025) : The Effect of Age Diversity in Groups on Peer Evaluations and Individual Performance, Kyklos, ISSN 1467-6435, Wiley, Hoboken, NJ, Vol. 79, Iss. 1, pp. 218-231,
https://doi.org/10.1111/kykl.70024

This Version is available at:
https://hdl.handle.net/10419/335600

Kyklos

WILEY

KYKLOS

ORIGINAL ARTICLE OPEN ACCESS

The Effect of Age Diversity in Groups on Peer Evaluations and Individual Performance

Katja Görlitz[1] | Tim Sels[2]

[1]HdBA, Hochschule der Bundesagentur für Arbeit, Mannheim, Germany | [2]FU Berlin, Berlin, Germany

Correspondence: Katja Görlitz (katja.goerlitz@hdba.de)

Received: 25 August 2024 | **Revised:** 10 September 2025 | **Accepted:** 23 October 2025

Keywords: Age diversity | expert evaluation | peer evaluation | performance

ABSTRACT

This study analyzes how individuals evaluate their peers' performance in a high-stakes tournament in response to being randomly assigned to an age homogenous or heterogenous group using data from two TV shows. The data also allow us to explore expert evaluations because it contains objective ratings from an independent expert. Additionally, this study investigates how age-diverse groups affect individual performance in professional golf tournaments. The results show that peer and expert evaluations as well as individual performance are lower in age-diverse groups. Further evidence suggests that these effects occur when group members are unfamiliar but fade away once group members have gotten to know each other.

JEL Classification: J14, M12, M54

1 | Introduction

The workforce is becoming increasingly diverse (Becker 2016). Population aging, increased longevity, and extensions of working life contribute to rising age diversity within organizations. Understanding how to compose effective work groups is a key challenge for managers, supervisors, and executive staff, as team composition directly affects productivity and decision-making. The previous empirical literature provides ambiguous findings on the relationship between demographic diversity such as gender (Joshi and Roh 2009), age or culture (Stahl et al. 2005), and group performance. While some studies find positive associations, highlighting improved creativity (e.g., Bantel and Jackson 1989) or problem-solving (e.g., Kilduff et al. 2000) in age-diverse teams, others report negative or null effects (e.g., Bunderson and Sutcliffe 2002, Timmerman 2000). A common view is that the context in which age diversity operates, such as task complexity, the degree of group interdependence, or the maturity of team relationships, plays a pivotal role in shaping its outcomes (Kunze et al. 2011; Williams and O'Reilly 1998). Nevertheless, many earlier studies offer limited insights into the question of whether potential performance differences can spill over to peer and expert evaluations.

This study adds novel insights into the evaluation behavior by analyzing how age diversity affects it in competitive, short-lived group settings. We investigate data from more than 2600 candidates from the TV cooking show "Come dine with me" where the aim is to prepare the most delicious three-course dinner and more than 1600 candidates from the TV show "Shopping Queen" where the candidates buy a stylish outfit in a given time. In both shows, five group members "cook" and "shop" against each other to win the high-stakes tournament prize. Importantly, the candidate with the highest scores given by the other players wins the prize. This is an interesting setting because workers face a trade-off. They could win with a higher likelihood by downgrading the other players' performance. Or they behave in a more fair and pro-social manner by reporting their correct subjective evaluations (which could be biased as well by preferences or tastes but not necessarily toward one's self-interest to

We acknowledge comments and help from Natalia Danzer, Olga Levina and Jan Marcus.

win). We investigate whether group members that are randomly assigned to either an age homogeneous or an age heterogenous group differ in how many points they award to their contestants.

This setting is also relevant from a managerial perspective. In many workplaces, direct supervision of employee output is not feasible, especially in team-based or multitasking environments. In such cases, large firms often rely on 360° feedback systems, where employees are assessed by multiple sources including peers, subordinates and supervisors (Edwards and Ewen (1996), Baroda et al. (2012), Mone and London (2018), and Rose and Biringer (2020) estimate a utilization rate of more than 90%). Although these systems aim to provide a more holistic view of performance, they may also be susceptible to interpersonal biases. This study contributes to this literature by analyzing the effects of peer evaluations. Because one of the two TV shows also reports the evaluations of an independent expert whose task is to report an objective review and who therefore does not compete in the tournament, we can also analyze how age diversity affects expert evaluations. This allows us to test whether age diversity affects third-party evaluations, which is critical in performance review contexts. To our knowledge, we are not aware of a study that investigates whether age diversity in groups affects the quality of peer and expert evaluations.

To examine whether these evaluations reflect differences in individual performance, we turn to professional golf tournament data from the PGA Tour. This setting is well-suited for us, because the tournament organizers assign the players randomly to groups, even though only individual performance determines the winner of the tournament. This is the same setting as in the TV shows. It differs from most of the previous literature as we do not analyze group performance in our study, but individual performance in response to changing the diversity composition of the group. This has practical relevance for common working situations where individuals perform individual tasks, but companies assign them to departments or divisions solely because of organizational reasons and not because working in groups is necessary to produce the output. Our analysis opens the black box of whether and how age diversity of groups affects individual behavior in these settings.

Our results indicate that individuals in age heterogenous groups award significantly fewer points to their peers. This is true regardless of using data from the cooking or the shopping show, meaning two settings that differ in the tasks that individuals have to perform and evaluate. This result spills over to the evaluation of an independent expert, suggesting that age diversity also affects the expert evaluations negatively. Investigating the golf data reveals that the reason for this evaluation differential is likely due to individual performance that is found to be lower in age-diverse groups as well. Further analysis shows that these results apply to settings where individuals have recently gotten to know each other. Therefore, they probably represent initial effects because further suggestive evidence shows that once group members have gotten familiar with one another, the performance effect vanishes.

The paper is organized as follows. The next section develops the hypotheses based on established theories and by integrating results from the empirical literature on age, demographic diversity, and evaluation biases. Section 3 presents the data, the empirical strategy, and the results on the relationship between age diversity and performance evaluations. Section 4 addresses how age diversity affects individual performance. The last section concludes, discusses limitations, and outlines directions for future research.

2 | Theoretical Background and Literature Review

2.1 | The Effect of Age Diversity on Peer and Expert Evaluations

Individuals seek to identify themselves as a member of a group based on self-categorization to gain social identity (Tajfel and Turner 1986; Turner 1987). The similarity-attraction theory states that groups have a stronger social cohesion if group members' attributes such as demographic characteristics like individuals' age are similar within a group (Berscheid and Walster 1969; Byrne 1971). This, in turn, can influence interpersonal dynamics such as cooperation, competition, and evaluation behavior. For example, Kelly and Presslee (2017) find that individuals in groups with stronger identification compete less and are, thus, less willing to win against another.

In newly formed groups, visible characteristics like age are among the first cues used for categorization (Fiske and Neuberg 1990; Harrison et al. 1998). While group cohesion and belonging may take time to develop through deeper knowledge of attitudes and values, surface-level cues dominate early impressions (Brewer and Harasty Feinstein 1999; Harrison et al. 2002). Within these early stages, group similarity may reduce competitive behavior due to stronger social identification and other-regarding preferences (Fehr and Schmidt 1999; Bolton and Ockenfels 2000; Charness and Rabin 2002).[1] For instance, Chen and Li (2009) demonstrate that individuals act more pro-socially toward those they perceive as part of their in-group. Conversely, age diversity may weaken these bonds and lead to less favorable evaluations.

This reasoning aligns with findings that age-diverse groups tend to experience weaker communication (Zenger and Lawrence 1989; Ellwart et al. 2014; De Meulenaere and Kunze 2020), higher interpersonal conflict (Jehn et al. 1997; Knight et al. 1999; Pelled et al. 2001; Colquitt et al. 2002; Luksyte et al. 2022), and lower group cohesion in terms of higher turnover of its members (Wagner et al. 1984; O'Reilly et al. 1989; Jackson et al. 1991; Tsui et al. 1992; Wiersema and Bird 1993; Kunze et al. 2021). Collectively, these factors (i.e., reduced empathy, diminished affiliation, and elevated conflict) can contribute to what we refer to as a "worse atmosphere," which can be characterized by interpersonal tension, lower psychological safety, and less trust. In performance contexts where peer evaluation influences outcomes, such dynamics may prompt individuals to rate others more harshly and/or to perform worse. We therefore propose the following:

Hypothesis 1. *Age diversity leads to lower peer evaluations among the group members.*

Beyond peer interactions, third-party observers may also influence evaluations. Independent evaluators, such as supervisors

or external judges, are assumed to be impartial, yet prior research suggests that evaluations can still reflect implicit biases (Murphy 2008; Scullen et al. 2000). Prendergast and Topel (1996) argue that even supervisors with no personal stake in outcomes may favor individuals on personal preferences. Bandiera et al. (2009) demonstrate that supervisors, who receive a fixed compensation that is independent of group performance, prefer those individuals to whom they are socially connected when choosing the members of their team. This may also be reflected in performance evaluations. In their studies, Stauffer and Buckley (2005) and Eren (2023) demonstrate the existence of racial and gender bias in supervisor assessments. However, in our setting, the expert evaluator does not interact with the contestants and provides ratings based solely on pre-recorded performance material. This suggests also an alternative and, for our purpose, more relevant explanation. Lower expert evaluations in age-diverse groups may reflect actual reductions in performance rather than evaluation biases. If age diversity negatively affects group dynamics and individual motivation, as we argue, this could manifest in objectively weaker performances that are then reflected in expert assessments. Thus, the second hypothesis is:

Hypothesis 2. *Age diversity among the group members leads to lower expert evaluations.*

2.2 | The Effect of Age Diversity on Individual Performance

Research results on the performance consequences of age diversity are contested. When individuals within a diverse group suffer from a worse atmosphere, this could spill over to a negative performance effect. While several empirical studies confirm that age diversity has indeed a negative impact on performance including slower decision-making, increased turnover, and coordination difficulties (Zajac et al. 1991; West et al. 1999; Timmerman 2000; Ely 2004; Leonard et al. 2004; Kearney and Gebert 2009; Kunze et al. 2011; Hafsi and Turgut 2013; Ali et al. 2014; De Meulenaere et al. 2016; De Meulenaere and Kunze 2020; Kunze et al. 2021; Luksyte et al. 2022), others find positive (Kilduff et al. 2000, Wegge et al. 2008, Kearney et al. 2009, Li et al. 2021) or null effects (Bantel and Jackson 1989; Wiersema and Bantel 1992; Simons et al. 1999; Bunderson and Sutcliffe 2002; van der Vegt and Bunderson 2005). These inconsistencies are highlighted in numerous meta-analyses and reviews, which emphasize that the impact of diversity is contingent on the task type, interaction frequency, and group maturity (Milliken and Martins 1996; Williams and O'Reilly 1998; Jackson et al. 2003; Joshi and Roh 2009; Wegge and Schmidt 2009; Bell et al. 2011; Boehm et al. 2011; van Dijk et al. 2012; Schneid et al. 2016).

Literature finding positive or insignificant effects investigates performance tasks that benefit from combining complementary knowledge and perspectives such as strategic decision-making, product development, interdisciplinary collaborations, or innovative solution findings (Bantel and Jackson 1989; Wiersema and Bantel 1992; Kearney et al. 2009; Luksyte et al. 2022). Also complex problem tasks that are best solved by group members with knowledge in different areas of expertise benefit from age diversity (Simons et al. 1999; Kilduff et al. 2000; Bunderson and Sutcliffe 2002; van der Vegt and Bunderson 2005; Wegge et al. 2008; Li et al. 2021). In contrast, settings where coordination is minimal and tasks are individual in nature may not harness the potential benefits of demographic diversity. Instead, the presence of unfamiliar or dissimilar group members may create subtle forms of psychological discomfort or distraction. Because our study analyzes data where performance is purely individual (even though individuals are part of groups) and because the task requires no innovative solution or knowledge in different areas of expertise, we hypothesize that we will find no positive effect. If age diversity reduces performance in this setting, it suggests that the social context alone without any functional interdependence can influence effort or concentration. The third hypothesis is:

Hypothesis 3. *Members of age-diverse groups exhibit a lower individual performance in our setting.*

A crucial theoretical distinction is that the negative effects of demographic diversity are likely strongest, when individuals have not yet developed familiarity or trust. Individuals divide themselves into social categories (Tajfel et al. 1971; Tajfel and Turner 1986; Turner 1987). These categories include among other things demographics (e.g., age and gender) as well as attitudes, beliefs, and values (Mannix and Neale 2005). While group members can immediately recognize some visible categories (age, gender), differences in others like attitudes, beliefs, and values only become apparent over time (Harrison et al. 1998; Harrison et al. 2002). Fiske and Neuberg's (1990) continuum model states that the first impression automatically or unconsciously leads to categorization based on demographics (age, gender) or other visible characteristics. After getting to know each other, other categories become more important. The same prediction arises from the dual process model stating that individuals choose the simpler information processing (Brewer and Harasty Feinstein 1999). Allport's (1954) contact-hypothesis states that getting to know each other more closely can reduce categorical prejudices. Our empirical settings are well-suited to examine this dynamic. In both TV shows and in the golf data, group members have minimal prior familiarity. This allows us to capture the initial impact of age diversity on evaluation and performance behavior. In certain constellations of the golf tournament, the PGA player groupings were already familiar with one another which we will exploit to test our fourth hypothesis which is:

Hypothesis 4. *Members of age-diverse groups exhibit a lower performance in our setting, particularly when they do not yet know each other.*

3 | The Effect of Age Diversity on Peer and Expert Evaluations

3.1 | Data and Empirical Model

3.1.1 | Data

This section uses data from two different sources. First, it exploits data from the German TV show "Das perfekte Dinner" based on the concept of the British TV series "Come Dine with

TABLE 1 | Summary statistics of the cooking and the shopping show.

	Cooking show				Shopping show			
Variables	**Obs.**	**Mean**	**Min**	**Max**	**Obs.**	**Mean**	**Min**	**Max**
Peer evaluation	10,580	7.64	0	10	6440	7.83	0	10
		(1.31)				(1.25)		
Expert evaluation		Not available			1610	7.46	1	10
						(1.26)		
Indicator for men (1—yes; 0—no)	2645	0.47	0	1	1610	0.001	0	1
Number of men per group	529	2.37	0	4	322	0.01	0	1
Age of the contestants in years	2645	39.66	18	85	1610	36.50	18	81
		(11.68)				(11.85)		
Standard deviation of the age in years per group	529	11.34	2.2	24.7	322	12.52	3.9	23.2

Note: The standard deviations of the means are shown in parentheses.

Me." Each calendar week, five volunteer contestants from one city compete against each other by cooking a three-course dinner of their own choice for the others. Every contestant serves a dinner in his or her apartment on another day of the workweek. The competitors do not know each other upfront and meet for the first time on Monday, which is the day when the first player prepares his or her dinner. All competitors aim to create the best dinner to win the cash prize of 3000 euro. The contestants determine the winner by peer evaluations. After the respective dinner evening, each of the four competitors anonymously awards scores in private on a scale from 0 to 10 to the chef. Thus, every chef can reach a maximum score of 40 points. The announcement of the winner takes place at the award ceremony at the end of the week after the last chef has served his or her dinner. While each contestant has the information about his or her overall score at the ceremony, the individual ratings for each candidate remain undisclosed and are not announced until the TV broadcasting. Because the scoring is recorded on camera and later broadcast to the public, a delayed social image concern may arise, as contestants know their behavior will eventually be visible to both peers and viewers. However, since scores are not revealed during the active competition, the immediate incentive structure is primarily shaped by an in-game strategy, interpersonal dynamics, and performance, rather than by post hoc reputation management. The data include TV episodes covering the period from January 2007 through January 2021 and encompasses 529 weeks with 2645 candidates as well as 10,580 peer evaluations.[2]

Second, this section uses data from the TV show *Shopping Queen* where five voluntary competitors compete in each calendar week by going shopping for at most 4 h to find a stylish outfit that is in line with a prescribed theme. The budget limit for the outfit is 500 euros, including clothing, shoes, accessories, hairstyle, and make-up. Each contestant must find, buy, and present the outfit on a different day during the working week. The candidates receive the prescribed theme on the day when all players meet each other for the first time. The other four competitors evaluate the performance anonymously and in private on a scale from 0 to 10 after the presentation of the outfit. In contrast to the cooking show, an expert rating also exists. At the end of the week, the German fashion expert Guido Maria Kretschmer always gives a score for each of the five candidates using the same scale from 0 to 10. Thus, the external expert rating influences the outcome of the competition. The expert evaluation of the candidates is based on the video recordings made. The critique and the awarded scores of the other candidates remain unknown to the expert. The expert gives his rating to each candidate at the award ceremony and announces the winner as well as the overall scores, while the individual ratings remain anonymous until the TV broadcasting. The maximum overall score is 50 points. Only the participant with the highest score will win a cash prize of 1000 euros.[3] The data covers 322 weeks from April 2013 through March 2021. In total, 1610 candidates, 6440 peer evaluations and 1610 expert evaluations are available.[4]

Table 1 contains the summary statistics for both data sources. The average of the peer evaluations given by the other players is 7.6 in the cooking and 7.8 in the shopping show. In the shopping show, the expert evaluation has a slightly lower mean of 7.5, but a similar standard deviation as the peer evaluations. While almost equal numbers of men and women are in the cooking show, the participants in the shopping show are almost exclusively women; because of the 1610 players, there were only two men. The chefs are older than the shopping contestants, and their standard deviation is slightly lower. We generated the standard deviation of the age of the contestants at the group level in order to give a more detailed view on the extent of age diversity in both data sources. As before, the age variation of the contestants is slightly lower in the cooking show. Nevertheless, both data sources show that age diversity varies greatly over the groups. While some groups are age homogenous (with a minimum value of 2.2 in the cooking show and 3.9 in the shopping show), others are rather heterogenous

(where the maximum is as high as 24.7 and 23.2, respectively). In particular, the group with the lowest overall variation consists of players having the following ages: 20, 23, 25, 25, and 25 years. In contrast, the most heterogenous group comprises players being 23, 32, 53, 63, and 85 years old.

3.1.2 | Empirical Model

The following OLS regression analysis estimates the effect of age diversity on the evaluation scores that each contestant *i* gets from each other player *j* in group (or equivalently week) *k*:

$$evaluation\ score_{i,j,k} = \alpha + \beta\ age\ diversity_k + X_i'\delta + X_j'\gamma + X_k'\eta + \varepsilon_{i,j,k} \quad (1)$$

where the dependent variable *evaluation score* refers to the peer evaluations. The independent variable of interest is *age diversity*. The regression controls for the characteristics of the player (X_i), the competitors (X_j), and group-specific covariates (X_k). The characteristics cover the age of the contestant and the age of the other players that account for potential correlations between age diversity and the players' ages. In the cooking show data, the corresponding regression can additionally incorporate the gender of both. The group-specific covariates include weekday fixed effects that control for the day of the week where a contestant cooks or shops absorb day-related scoring effects.[5] They also include the average age per group and for the cooking show additionally the share of men per group. ε is the idiosyncratic error term. Inference is based on heteroskedasticity-robust standard errors.

The identification strategy exploits the exogenous variation in age diversity across groups to estimate the effect on the players' evaluation score. The variation derives from the setting of the TV show. The TV production team assigns the contestants to their competitors based on geographic availability and logistical constraints,[6] who could thereby not self-select themselves into another group. As already explained in the data section, this is because the locations of the series change from week to week to another city where the contestants have to have an apartment. While we cannot find evidence of demographic engineering for "good TV," this is not a perfect randomized controlled trial, but the allocation process introduces substantial exogenous variation in group age composition. In addition, the competitors meet for the first time on the day when the recording of the first episode begins and the producers confirm they do not disclose peer group details to contestants in advance.

To find out whether age diversity only affects the peer evaluations or also spills over to the evaluation of an independent expert, the model estimates Equation (1) again using the evaluation score given by the expert from the shopping show as the dependent variable. Apart from that, this specification also changes the control variables. As in the analysis of the shopping data before, the model cannot take gender into account because almost all players are women and the expert is always a man. In the sensitivity analysis, the age of the expert is taken into consideration to account for the fact that the aging of the expert "Guido Maria Kretschmer" may lead to changes in preferences and tastes over time, which could potentially affect his evaluation behavior.

Because the previous literature analyzes alternative definitions of age diversity (Harrison and Klein 2007), we define *age diversity* in different ways to prove the robustness of our results. The most straightforward measure is to use the standard deviation of the contestants' age per group that varies tremendously across groups (see again Table 1). This measure considers absolute age differences. To analyze relative differences, we additionally define age diversity at the group level as the coefficient of variation which refers to the standard deviation divided by the average age (Timmerman 2000) as well as the standard deviation of the logarithm of participants' age (Leonard et al. 2004). These measures shed light on the questions of whether differences at older ages might be less pronounced than at younger ages. Furthermore, we allow for a more flexible functional form by estimating quartiles and tertiles of the group-specific standard deviation of age. This means estimating the standard deviation, sorting them in descending order and dividing them into four or three groups of equal size, respectively. Put differently, the lowest (highest) quartile contains the groups that belong to the bottom (top) 25% of all groups being characterized by the lowest (highest) age diversity. By doing so, this specification analyzes whether the effect is linear over age diversity or whether some parts of the distribution are affected more severely. Additionally, we present results from a fractionalization index that captures the likelihood that two randomly selected group members fall into different age categories. This index reflects categorical age heterogeneity rather than continuous variation, providing an alternative lens on group diversity.[7]

Further sensitivity analyses include the estimation of alternative standard errors. Because the players interact with each other at the group level, their errors might be serially correlated (Moulton 1990). Without accounting for the correlated errors, the models would underestimate standard errors and even nonsignificant effects could become statistically significant. Therefore, further sensitivity specifications cluster the standard errors at the group level, i.e., using week–year clusters and bootstrapped standard errors. This specification is not meaningful for the expert evaluation because only one person evaluates others, ruling any group interrelation out that Moulton (1990) describes. Re-estimating the main results by applying the ordered Logit and the ordered Probit model serves as a sensitivity check to find out whether the linear probability model is the adequate model specification. Furthermore, Schüller et al. (2014) demonstrate that the age difference between the player and the rater affects the performance ratings, which is why we introduce the age difference as an additional control variable.

3.2 | Results

Table 2 shows that age diversity reduces the peer evaluation scores that a player gets. Column (1) presents the results from

TABLE 2 | Results of age diversity on peer and expert evaluation.

	Peer evaluation		Expert evaluation
	Cooking show (1)	**Shopping show (2)**	**Shopping show (3)**
Age diversity defined as ...			
The standard deviation of the age by group	−0.021***	−0.026***	−0.024*
	(0.004)	(0.006)	(0.012)
The coefficient of variation: standard deviation divided by the average age by group	−0.843***	−1.010***	−0.975**
	(0.151)	(0.236)	(0.468)
The standard deviation of the logarithm of the age by group	−0.953***	−1.611***	−1.087**
	(0.163)	(0.275)	(0.535)
Quartiles of the standard deviation of the age by group			
First (lower) quartile		*Reference group*	
Second quartile	−0.021	−0.057	−0.009
	(0.036)	(0.045)	(0.087)
Third quartile	−0.168***	−0.055	−0.148
	(0.037)	(0.044)	(0.091)
Fourth (upper) quartile	−0.194***	−0.268***	−0.205**
	(0.037)	(0.052)	(0.102)
Tertiles of the standard deviation of the age by group			
First (lower) tertile		*Reference group*	
Second tertile	−0.109***	−0.077**	−0.002
	(0.031)	(0.038)	(0.075)
Third (upper) tertile	−0.182***	−0.238***	−0.201**
	(0.032)	(0.042)	(0.090)
Fractionalized index for the age in bins by group	−0.293***	−0.772***	−0.941***
	(0.095)	(0. 135)	(0.250)
Age of the peer who awards score	Yes	Yes	No
Age of player being evaluated	Yes	Yes	Yes
Gender of the peer who awards score	Yes	No	No
Gender of player being evaluated	Yes	No	No
Weekday fixed effects	Yes	Yes	Yes
Average age per group	Yes	Yes	Yes
Share of men per group	Yes	No	No
Observations	10,580	6440	1610

Note: The table reports the OLS results from regressing the peer evaluation (from the cooking and the shopping show in Columns (1) and (2), respectively) and from regressing the expert evaluation in the shopping show (Column (3)) on various measures of age diversity in addition to further covariates. See Equation (1) for further information. Statistical significance: $p < 0.1$*, $p < 0.05$**, and $p < 0.01$***.

the cooking show, documenting that an increase in the standard deviation by one unit reduces the evaluation scores given by each peer by −0.021. Related to the fact that always four players evaluate the remaining contestant, the overall score is 0.084 points lower. As the variation between the most homogenous and the most heterogenous group is 22.5 (=24.7—2.2) as can be seen from Table 1, this relates to an overall penalty of the most heterogenous group compared to the most homogeneous group of almost two points (=22.5x − 0.084). This is a nonnegligible effect size. In the shopping show, the corresponding estimate of the age diversity is −0.026 points for every player (Column (2)). The finding that age diversity decreases peer evaluations is robust to

analyzing alternative measures of age diversity. Interpreting the results of the quartiles and the tertiles further reveals that the effect of age diversity is slightly more pronounced when comparing groups with the lowest compared to groups with the highest age diversity.

The third column reveals that age diversity even influences the evaluation scores given by an independent expert who does not benefit himself from up- or downgrading others. In every one of the used definitions of age diversity, the expert gives on average a lower evaluation score to players from more age-diverse groups. When investigating quartiles or tertiles, groups being on the top of the diversity distribution experience the largest deduction of the expert evaluation. Nevertheless, the estimates of the expert evaluation are smaller in magnitude compared to the peer evaluations and lack statistical precision.

Table A1 contains the results of the sensitivity analyses. The sensitivity checks for the peer evaluation demonstrate that the results are robust to clustering and bootstrapping the standard errors, different estimation models (ordered Logit, ordered Probit) and using the age difference as further control variable. The same conclusion applies to analyzing expert evaluations, except for running ordered logit and probit specifications. Further robustness checks reinforce our main conclusion when adding the experts' age and its square as control variables to capture a more flexible nonlinear functional form. Thus, we can rule out the fact that the experts ages over time influences his scoring behavior by gradual shifts in preferences or standards. In conclusion, peers understate their contestants' performance in age-diverse groups in high-stakes tournaments. This reinforces our first hypothesis. There is also evidence for the second hypothesis, even though these estimates are less precise compared to the estimates of the peer evaluations.

4 | The Effect of Age Diversity on Individual Performance

4.1 | Data and Empirical Model

4.1.1 | Data

This section uses PGA Golf Tour 2002 data from Guryan et al. (2009). The PGA Golf Tour organizes various tournaments with professional golfers mainly in the United States. Each tournament has four rounds. The organizer assigns the golfers randomly into groups of three people which remain unchanged until the end of the second round. This is why this study analyzes the first and second rounds only to test our third hypothesis on performance effects. One concern when comparing the golf data with the TV show data is that the number of players per group differs. Although group size could potentially influence interaction dynamics, Laughlin et al. (2006) show that groups of three, four, and five people perform similarly in a complex group task. To test our fourth hypothesis, we exploit a particular feature of the data. When the organizers assign the golfers to the groups, they consider the players' performance by creating groups of golfers who belong to the same performance category and remain unchanged during the season. Category 1 refers to the best and Category 3 to the worst players. We expect that members of groups with Category 3 players do not previously know each other. This is because these groups consist of participants such as local qualifiers who have participated few times (5 years) on the PGA Golf Tour and played few tournaments (3.9) within the PGA Golf Tour. Furthermore, Category 3 includes approximately 30% more players compared to Category 1. For this reason, we refer to Category 3 players as "previously unknown group members." Category 1 players include tournament winners and the top 25 money earners from the previous year, as well as PGA Golf Tour life members like Tiger Woods who have achieved outstanding career achievements. We expect that Category 1 golfers more often previously know each other. On average, a Category 1 golfer has experienced 12 years on the PGA Golf Tour and participated in 12.6 tournaments. We refer to Category 1 players as "previously known group members" in the remaining study.

All players compete against each other, including the players within each group, to win the high-stakes prize. On average, a tournament allocates 3.7 million dollars in prize money of which the winner gets about 18% and the top 10 golfers approximately 60%. The golfer who attains the lowest cumulative number of strokes wins. Thus, player A with 70 strokes performs better than golfer B with 72 strokes. For ease of interpretation, we define a new variable, henceforth the performance score, which multiplies the number of strokes by −1. The higher the performance of a player is (because of having fewer strokes), the higher is the corresponding performance score. We merge data on the age of the golfers to the PGA Golf Tour 2002. The age data are primarily sourced from the following websites: Wikipedia, PGA, BlueGolf, ESPN, and Yahoo!Sports. Using data on the first and second rounds leaves us with 193 players from 992 groups and 2976 performance scores.

Table 3 presents the summary statistics for the PGA Golf Tour 2002 data separately for the previously unknown and the previously known group members. The previously unknown group members have the lowest and the previously known group members have the highest performance score because they require the most (72.2) and the fewest (70.5) strokes on average, respectively. This also corresponds to the handicap (ability score), which is better for known group members. Female golfers did not participate in the tournaments studied. The group of unknown members is younger than the group of known members. Generating the standard deviation of the age of the player at the group level provides a more detailed view on the extent of age diversity. Both groups show a wide range of age diversity. While some groups are age homogenous (with a minimum value of 0.6), others are more age heterogenous (where the maximum is as high as 17.8 and 15.3 in either of the two groups). The group with the lowest variation consists of players having, e.g., the following ages: 23, 24, and 24 or 39, 40, and 40 years. In contrast, the most heterogenous group of previously unknown members comprises players who are 24, 47, and 59 years old.

TABLE 3 | Summary statistics of the PGA Golf Tour 2002 data.

	Group members who are previously ...							
	Unknown				Known			
Variables	**Obs.**	**Mean**	**Min**	**Max**	**Obs.**	**Mean**	**Min**	**Max**
Performance score (negative number of strokes)	462	−72.24	−84	−63	2514	−70.48	−84	−61
		(3.51)				(3.08)		
Ability (handicap)	462	1.2	−1.5	8.1	2514	−0.3	−2.5	4.1
		(1.5)				(0.7)		
Age of the players in years	462	31.94	17	59	2514	36.47	22	59
		(7.44)				(5.88)		
Standard deviation of the age in years per group	154	6.37	0.6	17.8	838	5.31	0.6	15.3

Note: We defined the performance score as the number of strokes multiplied by -1. This guarantees that a better golfer with fewer strokes has a higher value on the performance score. The standard deviations of the means are shown in parentheses.

4.1.2 | Empirical Model

The following OLS regression analysis estimates the effect of age diversity on the performance scores that each player i achieves in group k:

$$performance\ score_{i,k} = \alpha + \beta\ age\ diversity_k + \delta X_i + X_k{}'\eta + \varepsilon_{i,k} \quad (2)$$

The dependent variable *performance score* refers to the players' golf performance, which is better the higher the score is. Following our analyses of peer evaluations in Section 3, the model uses various measures of *age diversity* as independent variable where the main specification uses the standard deviation of the players' age per group. The regression controls for the age of the player (X_i) to account for potential correlations between the age diversity and the players' ages. X_k comprises round fixed effects that absorb round-related scoring effects including weather conditions and it includes the average age of the group. ε is the idiosyncratic error term. Inference is based on heteroskedasticity-robust standard errors. Exploiting the fact that golfers are randomly assigned to groups, the identification strategy uses the exogenous group-specific variation in age diversity in the groups to estimate the effect on players' performance scores.

Further sensitivity analyses reveal whether age diversity affects the objective performance scores. Clustering the standard errors at the group level, which refers to group-round clusters and bootstrapping the standard errors examines whether alternative standard errors produce robust results. An ordered Logit and an ordered Probit model retest whether a nonlinear model can confirm the OLS results. In contrast to the dinner and the shopping show, further sensitivity analyses add control variables of players' ability (handicap) and experience (years on tour). To support the fourth hypothesis, another model specification examines first round effects only, because players have just recently gotten to know each other in the first round.

4.2 | Results

Table 4 summarizes the effects of age diversity on the golf performance score. Column (1) shows that age diversity has a significant negative effect on the performance of previously unknown group members. A one-unit increase in the standard deviation reduces performance by 0.17 additional strokes per round for each previously unknown group member. Because every golfer plays two rounds in the same group, they perform 0.34 strokes worse. The deviation between the most homogeneous and heterogenous groups is 17.2 (=17.8—0.6). Thus, the most age-diverse group is at a disadvantage of about 5.87 strokes. If one takes the respective tournament winner as a reference, a golfer will earn about $544,000 less in prize money. This conclusion is robust to analyzing alternative measures of age diversity. In contrast, the results of the previously known members in Column (2) are statistically insignificant regardless of using alternative measures of age diversity. Importantly, the magnitude of the coefficients is much smaller compared to the results of the unknown members. This indicates that the difference in the results from Columns 1 and 2 is not due to a lack of efficiency but shows that no performance effect for players with previously known members exists.

As another sensitivity check, we ensure that the observed effects are truly driven by group-level age diversity rather than by the individual ages of the participants; we also examined whether a player's own age directly affects performance. Our results indicate that slight performance penalties related to age are confined to groups where players are unfamiliar with one another. In these groups, older players tend to perform slightly worse as age increases. In contrast, in groups where players are more likely to know each other, individual age has no observable impact on performance. This suggests that familiarity may buffer any age-related performance pressures, reinforcing our broader finding that social context moderates the effects of age diversity. The sensitivity analyses shown in Table A2 reinforce our main results in all specifications. In conclusion, this confirms our third and fourth hypotheses.

TABLE 4 | Results of age diversity on the performance score.

	Performance score of members of previously	
	Unknown groups (1)	**Known groups (2)**
Age diversity defined as ...		
The standard deviation of the age by group	−0.171***	0.011
	(0.050)	(0.027)
The standard deviation divided by average age by group	−5.677***	0.598
	(1.649)	(0.962)
The standard deviation of the logarithm of the age by group	−5.374***	0.551
	(1.622)	(0.904)
Quartiles of the standard deviation of the age by group		
First (lower) quartile	Reference group	Reference group
Second quartile	−0.154	0.271
	(0.417)	(0.176)
Third quartile	−0.480	−0.073
	(0.471)	(0.173)
Fourth (upper) quartile	−1.178 **	0.143
	(0.485)	(0.181)
Tertiles of the standard deviation of the age by group		
First (lower) tertile	Reference group	Reference group
Second tertile	−0.064	0.199
	(0.383)	(0.148)
Third (upper) tertile	−1.384***	−0.018
	(0.425)	(0.157)
Fractionalized index for the age in bins by group	−2.193**	−0.465
	(0.868)	(0.323)
Age of player	Yes	Yes
Average age by group	Yes	Yes
Round fixed effects	Yes	Yes
Observations	462	2514

Note: The table shows the effects of the OLS regression of the performance score which is the negative number of a player's strokes using the PGA Golf Tour 2002 data. The independent variable of interest is based on different definitions of age diversity. The model includes further fixed effects. Column (1) refers to previously unknown group members and Column (2) focuses on previously known group members. Statistical significance: $p < 0.1$*, $p < 0.05$**, and $p < 0.01$***.

5 | Conclusion

This study contributes new insights into the effects of age diversity on peer and expert evaluations as well as on individual performance in small group settings where the performance does not require teamwork or diverse expert knowledge. Using data from two TV shows, the results show that age-diverse groups award significantly fewer points in peer evaluations compared to homogeneous groups. The same applies to the expert evaluations of independent experts who do not participate in the tournament themselves, albeit the estimates are less precise. The reason for the systematically lower peer and expert evaluations might stem from subjective biases due to a worse group atmosphere. Alternatively, the group atmosphere could affect the work effort which in turn could spill over to individual performance. Exploiting data from the PGA golf tour allowed us to show that age diversity in tournaments indeed spills over to individuals' performance when group members do not previously know each other. Once the group members have gotten to know each other, this performance effect renders statistically insignificant.

Our findings enrich the literature on demographic diversity in several ways. While existing research has frequently focused on how diversity shapes group-level outcomes, we demonstrate that age diversity can also produce individual-level disadvantages.

Even without direct teamwork, group composition matters, indicating a broader influence of social context on individual behavior. The observed negative effect on individual performance suggests that social context alone—without functional interdependence—can create psychological discomfort or distraction that undermines effort and concentration. This implies that the mere presence of dissimilar group members can impose a social cost, even when the typical task-related benefits of diversity are absent. At the same time, our findings do not contradict theories that highlight the benefits of diversity in complex, collaborative tasks (e.g., improved creativity or problem-solving). Rather, they emphasize that such benefits did not emerge in our settings, where interactions were short-lived and task interdependence was minimal. This reinforces the idea that the advantages of age diversity manifest when diverse expertise can be shared or when teams have time to develop synergy, whereas in immediate competition settings, the drawbacks dominate. Showing that the effect of age diversity is heterogenous depending on group familiarity integrates our results with existing theory, affirming classic social identity mechanisms in the short term. Altogether, this suggests that age diversity can be a liability in a broader range of organizational contexts than previously emphasized, particularly in the absence of active management or sufficient time for familiarity to develop.

Despite these contributions, this study is not without limitations. First, the group sizes differ across settings (five members in the TV shows vs. three in the PGA Tour), which might influence group dynamics, interpersonal interactions, and the magnitude of diversity effects. Second, while we focus on age diversity, other salient demographic dimensions such as gender, ethnicity or socioeconomic background could also influence group behavior. Third, our analyses capture short-term effects, especially in the golf data, which are limited to two rounds. We therefore cannot directly test when the effects of age diversity start to fade. Finally, we use distinct empirical settings: TV shows and professional sports. While this heterogeneity limits direct comparability, it also serves to analyze the broader effects of age diversity across contexts. Building on our findings and the aforementioned limitations, we identify several avenues for future research. First, the robustness of our results with regard to group size and exploiting other settings than TV shows or sports could be tested. Second, analyzing other demographic diversity dimensions could additionally contribute to the previous literature. Intersectional analyses examining interactions between age, gender, and ethnicity would offer a richer understanding of how demographic diversity shapes group dynamics. Third, longitudinal studies could assess how the effects of age diversity evolve as group members interact over time and familiarity increases.

Despite limitations and various questions that remain for future research, our findings contribute to a deeper understanding of how age diversity influences individual behavior in high-stakes, short-term group settings. These results have practical implications. They imply that organizations should approach demographic diversity with nuance. While diverse workforces can bring long-term benefits for innovation and decision-making in interdependent tasks, our study highlights short-run frictions in settings where demographic diversity is salient but irrelevant to task requirements. This is comparable to settings where the organizational structure divides individuals' groupwise, e.g., into sections or departments, but each worker performs his/her own tasks individually. Managers should be aware that peer and supervisor evaluations, including widely used 360° feedback systems, may inadvertently disadvantage individuals assigned to diverse teams. This could affect rewards, promotions, and motivation. Even though this disadvantage of age-diverse groups mirrors performance differences that could explain pay differentials, it is exogenous from an individual point of view and only represents the "luck" of being assigned to an age-homogeneous group. This suggests a need for careful team design, evaluation safeguards, and performance monitoring, particularly during the initial stages of team formation. Policies that foster familiarity and trust, such as team-building activities or longer term team assignments, could help to avoid early-stage drawbacks of demographic diversity.

Acknowledgments

Open Access funding enabled and organized by Projekt DEAL. [Correction added on 7 November 2025, after first online publication: Projekt DEAL funding statement has been added.]

Data Availability Statement

The data that support the findings of this study are available from Tim Sels upon reasonable request.

Endnotes

[1] See Cooper and Kagel (2006) and Fehr and Schmidt (2006) for a review of this literature.

[2] Some episodes are excluded, e.g., where only four contestants participated in the show (e.g., in calendar weeks with a holiday), where shows with other participants/rules took place (e.g., with celebrities as contestants) and where episodes or information like the candidates' age were unavailable.

[3] In rare cases, the winner receives only a noncash prize (e.g., a bag worth 4000 euros).

[4] Again, this study excludes shows with missing information and irregular shows like those with four competitors or special shows.

[5] Previous research shows that day-related effects can have a significant impact on the performance rating (Schüller et al. 2014; Blum and Wenskat 2020).

[6] Contestants are chosen based on their location (e.g., Berlin, Munich) and then grouped according to that week's filming schedule, not based on group composition, personality, or demographics. There is no evidence of demographic engineering for "good TV," and the producers confirm that they do not disclose peer group details to contestants in advance.

[7] To construct this index, we define age bins based on the following age distribution: Bin 1 for the youngest 25%, Bin 3 for the oldest 25%, and Bin 2 for all others. The specific cutoffs are as follows: 18–30, 31–48, 49+ for Come Dine with Me; 18–26, 27–45, 46+ for Shopping Queen; 17–31, 32–40, 41+ for the PGA Tour.

References

Ali, M., Y. L. Ng, and C. T. Kulik. 2014. "Board Age and Gender Diversity: A Test of Competing Linear and Curvilinear Predictions." *Journal of Business Ethics* 125, no. 3: 497–512.

Allport, G. W. 1954. *The Nature of Prejudice*. Addison-Wesley.

Bandiera, O., I. Barankay, and I. Rasul. 2009. "Social Connections and Incentives in the Workplace: Evidence From Personnel Data." *Econometrica* 77, no. 4: 1047–1094.

Bantel, K. A., and S. E. Jackson. 1989. "Top Management and Innovations in Banking: Does the Composition of the Top Team Make a Difference?" *Strategic Management Journal* 10, no. S1: 107–124.

Baroda, S., C. Sharma, and J. K. Bhatt. 2012. "360 Degree Feedback Appraisals—An Innovative Approach of Performance Management System." *International Journal of Management and Information Technology* 1, no. 2: 53–66.

Becker, F. 2016. *Teamarbeit, Teampsychologie, Teamentwicklung. So führen Sie Teams*. Springer-Verlag.

Bell, S. T., A. J. Villado, M. A. Lukasik, L. Belau, and A. L. Briggs. 2011. "Getting Specific About Demographic Diversity Variable and Team Performance Relationships: A Meta-Analysis." *Journal of Management* 37, no. 3: 709–743.

Berscheid, E., and E. H. Walster. 1969. *Interpersonal Attraction*. Addison-Wesley.

Blum, Peter, and Wenskat, Marc (2020), Why Monday Never Wins. An Example of the Secretary Problem, June.

Boehm, S. A., M. K. Baumgaertner, D. J. G. Dwertmann, and F. Kunze. 2011. "Age Diversity and Its Performance Implications—Analyzing a Major Future Workforce Trend." In *From Grey to Silver*, edited by S. Kunisch, S. Boehm, and M. Boppel, 123–144. Springer-Verlag.

Bolton, G. E., and A. Ockenfels. 2000. "ERC: A Theory of Equity, Reciprocity, and Competition." *American Economic Review* 90, no. 1: 166–193.

Brewer, M. B., and A. S. Harasty Feinstein. 1999. "Dual Processes in the Cognitive Representation of Person and Social Categories." In *Dual-Process Theories in Social Psychology*, edited by S. Chaiken and Y. Trope, 255–270. Guilford Press.

Bunderson, J. S., and K. M. Sutcliffe. 2002. "Comparing Alternative Conceptualizations of Functional Diversity in Management Teams: Process and Performance Effects." *Academy of Management Journal* 45, no. 5: 875–893.

Byrne, Donn (1971): The Attraction Paradigm, Academic Press.

Charness, G., and M. Rabin. 2002. "Understanding Social Preferences With Simple Tests." *Quarterly Journal of Economics* 117, no. 3: 817–869.

Chen, Y., and S. X. Li. 2009. "Group Identity and Social Preferences." *American Economic Review* 99, no. 1: 431–457.

Colquitt, J. A., R. A. Noe, and C. L. Jackson. 2002. "Justice in Teams: Antecedents and Consequences of Procedural Justice Climate." *Personnel Psychology* 55, no. 1: 83–109.

Cooper, D. J., and J. H. Kagel. 2006. "Other-Regarding Preferences." In *The Handbook of Experimental Economics, Volume 2*, edited by J. H. Kagel and A. E. Roth, 1st ed., 217–289. Princeton University Press.

De Meulenaere, K., C. Boone, and T. Buyl. 2016. "Unraveling the Impact of Workforce Age Diversity on Labor Productivity: The Moderating Role of Firm Size and Job Security." *Journal of Organizational Behavior* 37, no. 2: 193–212.

De Meulenaere, K., and F. Kunze. 2020. "Distance Matters! The Role of Employees' Age Distance on the Effects of Workforce Age Heterogeneity on Firm Performance." *Human Resource Management* 60, no. 4: 499–516.

Edwards, M. R., and A. J. Ewen. 1996. "How to Manage Performance and Pay With 360-Degree Feedback." *Compensation and Benefits Review* 28, no. 3: 41–46.

Ellwart, T., S. Bündgens, and O. Rack. 2014. "Managing Knowledge Exchange and Identification in Age Diverse Teams." *Journal of Managerial Psychology* 28, no. 7/8: 950–972.

Ely, R. J. 2004. "A Field Study of Group Diversity, Participation in Diversity Education Programs, and Performance." *Journal of Organizational Behavior* 25, no. 6: 755–780.

Eren, O. 2023. "Potential In-Group Bias at Work: Evidence From Performance Evaluations." *Journal of Economic Behavior & Organization* 206: 296–312.

Fehr, E., and K. M. Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics* 114, no. 3: 817–868.

Fehr, Ernst and Schmidt, Klaus M. (2006): The Economics of Fairness, Reciprocity and Altruism—Experimental Evidence and New Theories, in: Kolm, S.-C. and Ythier, J. M. (ed.), *Handbook of the Economics of Giving, Altruism and Reciprocity*, Volume 1, North Holland: w. p., 1st Edition, 615–691.

Fiske, S. T., and S. L. Neuberg. 1990. "A Continuum of Impression Formation, From Category-Based to Individuating Processes: Influences of Information and Motivation on Attention and Interpretation." In *Advances in Experimental Social Psychology, Volume 23*, edited by M. P. Zanna, 1–74. Academy Press.

Guryan, J., K. Kroft, and M. J. Notowidigdo. 2009. "Peer Effects in the Workplace: Evidence From Random Groupings in Professional Golf Tournaments." *American Economic Journal: Applied Economics* 1, no. 4: 34–68.

Hafsi, T., and G. Turgut. 2013. "Boardroom Diversity and Its Effect on Social Performance: Conceptualization and Empirical Evidence." *Journal of Business Ethics* 112, no. 3: 463–479.

Harrison, D. A., and K. J. Klein. 2007. "What's the Difference? Diversity Constructs as Separation, Variety, or Disparity in Organizations." *Academy of Management Review* 32, no. 4: 1199–1228.

Harrison, D. A., K. H. Price, and M. P. Bell. 1998. "Beyond Relational Demography: Time and the Effects of Surface- and Deep-Level Diversity on Work Group Cohesion." *Academy of Management Journal* 41, no. 1: 96–107.

Harrison, D. A., K. H. Price, J. H. Gavin, and A. T. Florey. 2002. "Time, Teams, and Task Performance: Changing Effects of Surface- and Deep-Level Diversity on Group Functioning." *Academy of Management Journal* 45, no. 5: 1029–1045.

Jackson, S. E., J. F. Brett, V. I. Sessa, D. M. Cooper, J. A. Julin, and K. Peyronnin. 1991. "Some Differences Make a Difference: Individual Dissimilarity and Group Heterogeneity as Correlates of Recruitment, Promotions, and Turnover." *Journal of Applied Psychology* 76, no. 5: 675–689.

Jackson, S. E., A. Joshi, and N. L. Erhardt. 2003. "Recent Research on Team and Organizational Diversity: Swot Analysis and Implications." *Journal of Management* 29, no. 6: 801–830.

Jehn, K. A., C. Chadwick, and S. M. B. Thatcher. 1997. "To Agree or Not to Agree: The Effects of Value Congruence, Individual Demographic Dissimilarity, and Conflict on Workgroup Outcomes." *International Journal of Conflict Management* 8, no. 4: 287–305.

Joshi, A., and H. Roh. 2009. "The Role of Context in Work Team Diversity Research: A Meta-Analytic Review." *Academy of Management Journal* 52, no. 3: 599–627.

Kearney, E., and D. Gebert. 2009. "Managing Diversity and Enhancing Team Outcomes: The Promise of Transformational Leadership." *Journal of Applied Psychology* 94, no. 1: 77–89.

Kearney, E., D. Gebert, and S. C. Voelpel. 2009. "When and How Diversity Benefits Teams: The Importance of Team Members' Need for Cognition." *Academy of Management Journal* 52, no. 3: 581–598.

Kelly, K., and A. Presslee. 2017. "Tournament Group Identity and Performance: The Moderating Effect of Winner Proportion." *Accounting, Organizations and Society* 56: 21–34.

Kilduff, M., R. Angelmar, and A. Mehra. 2000. "Top Management-Team Diversity and Firm Performance: Examining the Role of Cognitions." *Organization Science* 11, no. 1: 21–34.

Knight, D., C. L. Pearce, K. G. Smith, et al. 1999. "Top Management Team Diversity, Group Process, and Strategic Consensus." *Strategic Management Journal* 20, no. 5: 445–465.

Kunze, F., S. A. Boehm, and H. Bruch. 2011. "Age Diversity, Age Discrimination Climate and Performance Consequences—A Cross Organizational Study." *Journal of Organizational Behavior* 32, no. 2: 264–290.

Kunze, F., S. A. Boehm, and H. Bruch. 2021. "It Matters How Old We Feel in Organizations: Testing a Multilevel Model of Organizational Subjective-Age Diversity on Employee Outcomes." *Journal of Organizational Behavior* 42, no. 4: 448–463.

Laughlin, P. R., E. C. Hatch, J. S. Silver, and L. Boh. 2006. "Groups Perform Better Than the Best Individuals on Letters-to-Numbers Problems: Effects of Group Size." *Journal of Personality and Social Psychology* 90, no. 4: 644–651.

Leonard, J. S., D. I. Levine, and A. Joshi. 2004. "Do Birds of a Feather Store Together? The Effects on Performance of Employees' Similarity With One Another and With Customers." *Journal of Organizational Behavior* 25: 731–754.

Li, Y., Y. Gong, A. Burmeister, et al. 2021. "Leveraging Age Diversity for Organizational Performance: An Intellectual Capital Perspective." *Journal of Applied Psychology* 106, no. 1: 71–91.

Luksyte, A., D. R. Avery, S. K. Parker, Y. Wang, L. U. Johnson, and L. Crepeau. 2022. "Age Diversity in Teams: Examining the Impact of the Least Agreeable Member." *Journal of Organizational Behavior.*

Mannix, E., and M. A. Neale. 2005. "What Differences Make a Difference? The Promise and Reality of Diverse Teams in Organizations." *Psychological Science in the Public Interest* 6, no. 2: 31–55.

Milliken, F. J., and L. L. Martins. 1996. "Searching for Common Threads: Understanding the Multiple Effects of Diversity in Organizational Groups." *Academy of Management Review* 21, no. 2: 402–433.

Mone, Edward M. and London, Manuel (2018): Employee Engagement Through Effective Performance Management. A Practical Guide for Managers, Routledge Taylor & Francis Group: Second Edition.

Moulton, B. R. 1990. "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units." *Review of Economics and Statistics* 72: 334–338.

Murphy, K. R. 2008. "Explaining the Weak Relationship Between Job Performance and Ratings of Job Performance." *Industrial and Organizational Psychology* 1: 148–160.

O'Reilly, C. A., III, D. F. Caldwell, and W. P. Barnett. 1989. "Work Group Demography, Social Integration, and Turnover." *Administrative Science Quarterly* 34, no. 1: 21–37.

Pelled, L. H., K. R. Xin, and A. M. Weiss. 2001. "No es como mí: Relational demography and Conflict in a Mexican Production Facility." *Journal of Occupational and Organizational Psychology* 74, no. 1: 63–84.

Prendergast, C., and R. H. Topel. 1996. "Favoritism in Organizations." *Journal of Political Economy* 104, no. 5: 958–978.

Rose, Dale S. and Biringer, Jesse C. (2020): Current Practices in 360 Feedback, 3D Group, 6th Edition.

Schneid, M., R. Isidor, H. Steinmetz, and R. Kabst. 2016. "Age Diversity and Team Outcomes: A Quantitative Review." *Journal of Managerial Psychology* 31, no. 1: 2–17.

Schüller, D., H. Tauchmann, T. Upmann, and D. Weimar. 2014. "Pro-Social Behavior in the TV Show "Come Dine With Me": An Empirical Investigation." *Journal of Economic Psychology* 45: 44–55.

Scullen, S. E., M. K. Mount, and M. Goff. 2000. "Understanding the Latent Structure of Job Performance Ratings." *Journal of Applied Psychology* 85, no. 6: 956–970.

Simons, T., L. H. Pelled, and K. A. Smith. 1999. "Making Use of Difference: Diversity, Debate, and Decision Comprehensiveness in Top Management Teams." *Academy of Management Journal* 42, no. 6: 662–673.

Stahl, G. K., M. L. Maznevski, A. Voigt, and K. Jonsen. 2005. "Unraveling the Effects of Cultural Diversity in Teams: A Meta-Analysis of Research on Multicultural Work Groups." *Journal of International Business Studies* 41: 690–709.

Stauffer, J. M., and M. R. Buckley. 2005. "The Existence and Nature of Racial Bias in Supervisory Ratings." *Journal of Applied Psychology* 90, no. 3: 586–591.

Tajfel, H., M. G. Billig, R. P. Bundy, and C. Flament. 1971. "Social Categorization and Intergroup Behaviour." *European Journal of Social Psychology* 1, no. 2: 149–178.

Tajfel, Henri and Turner, John C. (1986): The Social Identity Theory of Intergroup Behavior, in Wochel, S. and Austin, W.G. (ed.), *Psychology of Intergroup Relations*, Nelson-Hall: 2nd Edition, pp. 7–24.

Timmerman, T. A. 2000. "Racial Diversity, Age Diversity, Interdependence, and Team Performance." *Small Group Research* 31, no. 5: 592–606.

Tsui, A. S., T. D. Egan, and C. A. O'Reilly III. 1992. "Being Different: Relational Demography and Organizational Attachment." *Administrative Science Quarterly* 37, no. 4: 549–579.

Turner, J. C. 1987. *Rediscovering the Social Group: A Social Categorization Theory.* Blackwell.

Van Der Vegt, G. S., and J. S. Bunderson. 2005. "Learning and Performance in Multidisciplinary Teams: The Importance of Collective Team Identification." *Academy of Management Journal* 48, no. 3: 532–547.

van Dijk, H., M. L. van Engen, and D. van Knippenberg. 2012. "Defying Conventional Wisdom: A Meta-Analytical Examination of the Differences Between Demographic and Job-Related Diversity Relationships With Performance." *Organizational Behavior and Human Decision Processes* 119, no. 1: 38–53.

Wagner, W. G., J. Pfeffer, and C. A. O'Reilly III. 1984. "Organizational Demography and Turnover in Top-Management Groups." *Administrative Science Quarterly* 29, no. 1: 74–92.

Wegge, J., C. Roth, B. Neubach, K.-H. Schmidt, and R. Kanfer. 2008. "Age and Gender Diversity as Determinants of Performance and Health in a Public Organization: The Role of Task Complexity and Group Size." *Journal of Applied Psychology* 93, no. 6: 1301–1313.

Wegge, J., and K.-H. Schmidt. 2009. "The Impact of Age Diversity in Teams on Group Performance, Innovation and Health." In *New Horizons in Management. Handbook of Managerial Behavior and Occupational Health*, edited by A.-S. G. Antoniou, C. L. Cooper, G. P. Chrousos, C. D. Spielberger, and M. W. Eysenck, 79–94. Edward Elgar Publishing.

West, Michael, Patterson, Malcom, Dawson, Jeremy and Nickell, Steve (1999), The Effectiveness of Top Management Groups in Manufacturing Organisations, Centre for Economic Performance, November.

Wiersema, M. F., and K. A. Bantel. 1992. "Top Management Team Demography and Corporate Strategic Change." *Academy of Management Journal* 35, no. 1: 91–121.

Wiersema, M. F., and A. Bird. 1993. "Organizational Demography in Japanese Firms: Group Heterogeneity, Individual Dissimilarity, and Top Management Team Turnover." *Academy of Management Journal* 36, no. 5: 996–1025.

Williams, K. Y., and C. A. O'Reilly III. 1998. "Demography and Diversity in Organizations: A Review of 40 Years of Research." *Research in Organizational Behavior* 20: 77–140.

Zajac, E. J., B. R. Golden, and S. M. Shortell. 1991. "New Organizational Forms for Enhancing Innovation: The Case of Internal Corporate Joint Ventures." *Management Science* 37, no. 2: 170–184.

Zenger, T. R., and B. S. Lawrence. 1989. "Organizational Demography: The Differential Effects of Age and Tenure Distributions on Technical Communication." *Academy of Management Journal* 32, no. 2: 353–376.

Appendix 1

TABLE A1 | Sensitivity checks of age diversity on peer and expert evaluation.

	Peer evaluation		Expert evaluation
	Cooking show	Shopping show	Shopping show
	(1)	(2)	(3)
Age diversity defined as the standard deviation of the age by group			
Model specification …			
Main results (taken from Table 2)	−0.021 ***	−0.026***	−0.024*
	(0.004)	(0.006)	(0.012)
Ordinary least squares (including 2nd polynomial of experts' age)			−0.021*
			(0.012)
Ordinary least squares (clustered standard errors)	−0.021***	−0.026*	
	(0.006)	(0.014)	
Ordinary least squares (bootstrapped standard errors)	−0.021***	−0.026*	
	(0.006)	(0.014)	
Ordered Logit (robust standard errors)	−0.028***	−0.039***	−0.022
	(0.005)	(0.009)	(0.018)
Ordered Probit (robust standard errors)	−0.016***	−0.022***	−0.016
	(0.003)	(0.005)	(0.010)
Controlling for the age difference	−0.013***	−0.023***	−0.043***
	(0.004)	(0.007)	(0.013)
Observations	10,580	6440	1610

Note: The table shows the results of the alternative OLS regressions, which measure again the peer evaluation (from the cooking and the shopping data in Columns (1) and (2), respectively) and the expert evaluation from the shopping show (third column). These specifications account for different standard errors, Logit and Probit models, and the experts' age. The last row adds the absolute age difference as a further control variable. Statistical significance: $p < 0.1$*, $p < 0.05$**, and $p < 0.01$***.

TABLE A2 | Sensitivity checks of age diversity on the performance score.

	Performance Score of members of previously	
	Unknown groups	**Known groups**
	(1)	**(2)**
Age diversity defined as the standard deviation of the age by group		
Model specification …		
Main results (taken from Table 4)	−0.171***	0.011
	(0.050)	(0.027)
Ordinary least squares (clustered standard errors)	−0.171***	0.011
	(0.057)	(0.030)
Ordinary least squares (bootstrapped standard errors)	−0.171***	0.011
	(0.059)	(0.031)
Ordered Logit (robust standard errors)	−0.096***	0.001
	(0.027)	(0.015)
Ordered Probit (robust standard errors)	−0.048***	0.003
	(0.015)	(0.009)
Controlling for ability and experience	−0.181***	0.031
	(0.048)	(0.027)
First round effects	−0.231***	0.015
	(0.070)	(0.034)

Note: The table presents the effect of age diversity on the PGA Golf Tour 2002 performance score for previously unknown (Column 1) and known group members (Column 2). The performance score is the negative number of strokes per round. The first row of Columns (1) and (2) shows the baseline specifications as specified in Equation (2). The other rows modify the model by estimating alternative standard errors, ordered Logit and Probit regressions or expanding the baseline regression by ability measured by the handicap and experience fixed effects measured by the years on tour. The observations are 462 (2514) for unknown (known) group members, respectively. The last row shows short-term effects by considering only the first round where observations are 240 for unknown and 1299 for known group members. Statistical significance: $p < 0.1$*, $p < 0.05$**, and $p < 0.01$***.