Behav Res (2011) 43:931–941
DOI 10.3758/s13428-011-0120-x

Testing the accuracy of the retrospective recall method used in expertise research

Robert W. Howard

Published online: 14 June 2011

Abstract Expertise typically develops slowly over years, and controlled experiments to study its development may be impractical. Researchers often use a correlational, retrospective recall method in which participants recall career data, sometimes over many years before. However, recall accuracy is uncertain. The present study investigated the accuracy of recalled career data for up to 38 years, in over 600 international chess players. Participants' estimates of their entry year into international chess, total career games played, and number of games in a typical year were compared with the known true values. Entry year typically was recalled fairly accurately, and accuracy did not diminish systematically with time since list entry from 10 years earlier to 25 or more years earlier. On average, games-count estimates were reasonably accurate. However, some participants were very inaccurate, and some were more inaccurate in their total-games counts and entry-year estimates. The retrospective recall method yields usable data but may have some accuracy problems. Possible remedies are outlined.

Keywords Retrospective recall method · Expertise · Autobiographical memory · Chess

Expertise may develop slowly over years, and controlled experiments to study its development may be impractical. Researchers often use a correlational, retrospective recall method in which participants estimate various career data. For instance, participants may estimate various career landmark dates (e.g., of domain entry, start of "serious practice," and peak performance level) or ages at which these occurred. Participants may estimate weekly hours of various types of practice since those dates (Charness, Tuffiash, Krampe, Reingold, & Vasyukova, 2005; Gobet & Campitelli, 2007). Total career practice hours then may be estimated from year-by-year practice estimates or from weekly practice hours in a typical year or in the past year.

Expertise studies using this method have had a great impact. Data from studies in music, chess, and sports have been used to suggest that practice alone is the key factor in determining expertise level (Ericsson, 2006; Howard, 2009), a view embraced in several disciplines and in the popular domain (Colvin, 2008; Gladwell, 2008; Ross, 2006).

Problems with the paradigm

However, various researchers have cited a major concern about the method (Cumming, Hall, & Starkes, 2005; Deakin, Cote, & Harvey, 2006): How accurate are such recalled career data? If estimates are very inaccurate, many study findings are questionable, and most evidence for the practice-alone view evaporates. Expertise researchers have acknowledged the problem. Charness et al. (2005, p. 162) stated "… we cannot ascertain the degree to which different skill groups or age groups in our study may have overestimated the frequency of duration of time spent on various chess activities."

In some expertise domains, the accuracy concern may not be so great. Careers may last only a few years, may begin at about the same time (e.g., the start of a school year), and practice may occur mostly in discrete, regularly scheduled sessions with others. But, in oft-used domains such as chess and music, careers can last decades, starting dates may be long ago, and practice may involve much self-scheduled, solitary, and sporadic study.

R. W. Howard (✉)
School of Education, University of New South Wales,
Sydney 2052, Australia
e-mail: rwh@unsw.edu.au

The problem is that autobiographical memory is notoriously unreliable (Bahrick, 1984; Burt, Kemp, & Conway, 2004; Williams, Conway, & Cohen, 2008). Recall may be distorted, and false memories can easily take root (Larsen & Conway, 1997; Niedzwienska, 2003). Sometimes recalled data can be checked against parental recollections, but then parental recall may also be distorted. Recall of events from many years before may involve reconstruction, may be biased by a personal narrative (e.g., "I am a talented wrestler and needed little practice to reach a high level"), and may be affected by various individual difference factors (Malmberg, 2007). Indeed, various studies of autobiographical memory have shown recall inaccuracy even with apparent "flashbulb" memories for highly significant occurrences (Christianson, 1989; Tekcan, Ece, Gulgoz, & Er, 2003; Wright, 1993). However, Potts, Belden, and Reese (2008) found reasonable accuracy of young adults' reports of childhood television viewing patterns when checked against their recall of details of programs. Campitelli, Parker, Head, and Gobet (2008) studied recall of chess game details up to 5 years previously of two highly skilled players and found reasonable recall accuracy.

Accuracy of recall of starting dates has not been a concern with researchers using the retrospective recall paradigm. However, such dates are an important component of total practice amount calculations, and claims about the importance of an early career start in attaining high performance levels rest on their accuracy. Accuracy in recalling historical dates has been examined by memory researchers. Some studies have shown reasonable accuracy in dating historical events over just the past few years (Friedman, 1993). Brown, Rips, and Shevell (1985) asked participants to estimate the dates of various historical events that had occurred since 1977. Participants on average were accurate to within half a year, but the retention interval was short. Kemp (1988) reported a "telescoping" effect; participants on average overestimated time since more recent events and underestimated time since more distant ones.

Diary studies of accuracy

Some researchers have tried to gauge the accuracy of recall of one type of career data, extent of practice, by diary studies. Participants may estimate weekly practice hours and then record actual practice hours in a diary for a week or longer. When mean diary values are compared with estimated values, a typical finding is that participants on average somewhat overestimate the number of practice hours (Deakin et al., 2006). For instance, Hodges and Starkes (1996), with a small sample of wrestlers, found an average diary value of around 11 h a week practice but an average estimate of around 17 h, and a sizeable correlation between typical week and diary week estimates. Ericsson, Krampe, and Tesch-Romer (1993) found that musicians overestimated practice hours by about 21%, and the correlation between diary and estimated hours was .74. Davids (2000) found that violinists overestimated their current average weekly practice time by 21.4%. De Bruin, Smits, Rikers, and Schmidt (2008) asked 36 chess players to fill out a practice diary for 3 weeks and correlated diary hours with their estimates for the previous year. For "serious study hours spent alone," the diary mean was 3.7 h and the weekly retrospective mean estimate was 4.38 h, and the correlation between the two was .6. For "serious chess play," the diary mean was 6.31 h and the estimate was 5.36 hours, and the correlation was .74. Deakin and Cobley (2003) directly observed 24 figure skaters' scheduled hours and actual practice hours and found that skaters overestimated practice hours.

Problems of diary studies

Diary studies have problems. First, such studies often use estimates for only the current year, and often focus on domains in which careers are short (Cumming et al., 2005). In domains in which careers may last many years, estimate accuracy is less certain. Second, the diary week may be unrepresentative of the entire year, and diary data themselves may be inaccurate. Indeed, because of problems with diary data, television ratings agencies shifted to "people meter" measures of viewing, where viewers hit a switch when starting and finishing viewing of a given program. But even such people-meter data have accuracy problems (Milavsky, 1992; Starkey, 2004). Third, sample sizes in diary studies often are too small to give much indication of variations in accuracy. Fourth, diary studies typically only examine the accuracy of practice estimates. How accurate are estimates of various career starting dates, such as that of "serious practice" or of peak performance level? For careers lasting just the past few years, such problems may be minor but may increase with career length.

A final problem with diary studies lies in their data analysis. Researchers typically use signed estimated values and present only grouped data, summing together over- and underestimates and getting an average value that may suggest better accuracy than is present in actuality. One remedy is to use individual deviation scores (the difference between the actual and estimated values), absolute values, and individual percentage differences. The present study mainly uses such absolute values. Signed values are given for comparison with diary studies.

Aims of the present study

The present study assessed the accuracy of recall of career data in international chess players, using an approach that overcomes some limitations of diary studies. Participants made estimates of three types of career data for which the true values are known. The first datum was their date of entry onto the international rating list (Howard, 2006), which could have occurred up to 38 years previously. List entry is an important career milestone and should be as significant as other career dates, such as that of starting serious practice or learning the chess moves. This date is not much discussed or thought about or asked about by tournament organizers, and neither would be dates of starting serious practice or learning the chess moves. The accuracy of estimating date of list entry should give some notion of the accuracy of estimates of dates of learning the moves and starting serious practice.

The second and third data points were number of internationally rated games in a typical year and career total of such games. Games usually last several hours and are dispersed over a career, which is roughly analogous to the way in which practice hours may be so dispersed. These data may give some notion of the accuracy of estimates of such dispersed practice hours, as discussed below. Participants estimated their total number of internationally rated games, and from their estimates of years in the domain and games in a typical year, another estimate of total number of games was calculated. The accuracies of two methods of estimating total games was compared. Finally, the sample was large, and some participants had been in the domain for decades, allowing for examination of questions of whether accuracy declines greatly with time in the domain and of individual variation in accuracy.

Expertise researchers have not been concerned greatly with these specific career data, but have typically asked about dates of learning the chess moves or of starting serious practice and estimates of numbers of practice hours. However, in some studies participants have estimated game data (Charness et al., 2005; Gobet & Campitelli, 2007). For most participants in expertise studies, objective checks of such estimates are not possible. The true values are unknown and mostly may be unknowable. But the date of entry into the international chess domain and the count of internationally rated games are known values. If participants are wildly inaccurate at recalling these data, then estimates of practice hours and starting dates may be just as inaccurate and may be unusable. If the estimates are accurate, researchers can have more confidence in estimates of other career data, such as practice hours. Even then, the problem of recalling the date of first international rating is closely related to that of recalling other career dates (e.g., of starting serious practice). If recall accuracy is poor, this would put in question the accuracy of other date estimates.

The data also may be of interest to researchers in autobiographical memory. Here is a rare case in which data recalled over many years can be compared with the known true values, with a large sample.

Method

Data sets

Two data sets were used.

International chess federation (FIDE) rating data The first FIDE rating list appeared in 1970. FIDE initially issued one list a year, then two (in January and July), in 2000 issued three, and until 2009 issued four. Howard (2006) created a database that contains all FIDE rating data. The rating lists include all players active in international tournaments since 1970 who achieve and keep a minimum performance level. Players in rated tournaments who lack an initial FIDE rating get on the FIDE list if they achieve and maintain that specified minimum performance level. However, FIDE is a federation of national chess associations, not of individuals. There are no joining fees or membership cards for individuals. Players who get a rating are not directly notified, and indeed some participants may not know their actual entry date. Until 1990, ratings were published only in issues of the periodical *Sahovsky Informator*. Rating lists nowadays are published on the FIDE website, which has an archive of separate lists from July 2001. Lists give a player's current performance rating on a scale from about 1200 to about 3000 (Elo, 1986). The numbers of internationally rated games in a rating period are tallied only from July 1985. The FIDE rating data are not perfect (Howard, 2006), and the data of the present participants were checked thoroughly. Rules to get on and stay on the rating list have changed. Males once needed a minimum rating of 2200, but FIDE has dropped the minimum periodically from 1993, and it now is 1200.

Survey data The second data set was gathered through a 23-question online survey that ran from June 2008 to January 2009. Online surveys allow for large, varied samples (Birnbaum, 2004). Professional translators translated questions into German, Russian, and Spanish, which with English are the major languages of chess players.

The survey was advertised as a study of the role of practice and natural talent in chess skill. Practice and views about natural talent were its main focus. Ads were posted in all chess newsgroups and as a news item on many chess sites, such as The Week in Chess, ChessBase, the FIDE

website, and some national federation websites. The ad stated that the survey was only for players who had a FIDE rating and that each participant needed to know his or her FIDE ID number, available on the FIDE website. Respondents were invited to go to a website that had four links labeled in the four languages, leading to the survey forms. No one was paid for participation. The ad stated that participants who requested it would be e-mailed a summary of results and that results would be posted on websites.

The author checked all responses, ensuring that respondents had a FIDE rating when they completed the survey and that ID numbers and database names matched. Players needed to look up their FIDE number and were asked when they had entered the rating list and how many FIDE rated games in total they had played. The "Submit" button only worked when the FIDE ID number box had been filled in. That was the only required field. Almost all participants inserted an e-mail address to get emailed results, and addresses generally were consistent with their identities. A response was eliminated if there was doubt, but such cases were few.

Only 77 responses were eliminated, most for lack of a FIDE rating. In these cases, the respondent inserted a code, but it and their name did not exist in the FIDE database. Two responses seemed frivolous, a few were less than 50% filled out, and some participants clicked the survey "Submit" button several times. Two players filled out the survey some months apart, an occasional problem with online surveys (Birnbaum, 2004). Most of their responses to the questions matched, and only the earlier completions were used. This left 680 usable completions, about 69.4% in English. When respondents specified a range (e.g., 4–6 games), the midpoint value was used (e.g., 5 games).

The relevant questions for the present study were:

1. Your FIDE ID number (FIDE code) is:
2. Your full name is:
3. In what year did you first appear on the FIDE rating list (e.g., first get an FIDE rating)?
4. About how many FIDE rated games a year have you played in a typical year since first appearing on the FIDE rating list? Please count only FIDE rated games and please answer just from memory.
5. About how many FIDE rated games have you played in total? Please answer just from memory.
6. If you would like to receive an emailed summary of the survey results, please insert your email address here.

Participants

These were 680 rated players, of whom 16 (2.35%) were female, and who were from many nationalities. Using FIDE's nation codes on the January 1, 2009, list, some of the percentages were as follows: Germany and England 9.1%, Spain 5.3%, and the United States 5.1%. However, some rated players live outside their country of origin. Participants entered the list at varying dates (some in 2008, and one in 1970). The median age on January 1, 2009, for the 619 participants with a known birth date was 35.67 years (SD = 12.15). Their mean peak rating by January 1, 2009, was 2151.43 (SD = 170.91). Eight were grandmasters. All were "elite" players, as anyone with a national or international rating is an elite player. However, skill varied from grandmaster to fairly ordinary club-player level, with peak ratings ranging from 2670 to 1518. About 17% of the participants had a peak rating below 2000. Not all participants completed all relevant survey questions, and the ns are given below. For the games-count computations, only the 629 participants who answered all of Questions 3, 4, and 5 and were first on the list from July 1985 (when FIDE first started reporting game totals) were used.

Procedure

Year of entering the FIDE rating list and FIDE-rated games counts were calculated from the FIDE database (Howard, 2006). An accuracy measure for entry year estimates was calculated by subtracting recalled year of entry from actual year of entry, excluding months. So, if a player actually entered the list any time in 2002 but estimated entry year as 2004, the measure was −2. If that player had estimated entry in 1998, the value would be 4. Zero showed an accurate estimate, 1 showed recalled entry a year before actual entry, and −1 showed recall entry a year later than actual entry. When the absolute value of the measure was greater than 5, the database again was checked thoroughly to see if the FIDE data were incorrect. For 2 participants, the FIDE data possibly were incorrect, so their entry year and games-count data were not used. All other participants so checked were retained in the study. Ten participants did not answer the year-of-entry question, and their data were not used.

Because many participants entered the list very recently, the recall task was easy. Nearly half had entered in the past 4 years. So, a supplementary analysis was done for the 262 participants who had entered in 2000 and earlier. This year was chosen because it gives a long time span, and participants could not have looked up their entry year on the FIDE website. The FIDE website has lists only back to 2001, and then one would have to trawl through many names and lists to find an entry date. Furthermore, to see if accuracy declines systematically with years since entry, players were placed into categories of years since entry date, with greatly differing sample sizes. These categories were as follows; 0–4 years from entry (n = 296), 5–9 years

(n = 166), 10–14 years (n = 105), 15–19 years (n = 56), 20–24 years (n = 20), and 25-plus years (n = 25). The accuracy measure was computed for each category.

Game totals in the database were computed up to the approximate time of taking the survey, which was between June 6, 2008, and January 4, 2009. Relevant FIDE lists came out on July 1, 2008, October 1, 2008, and January 1, 2009. Games counts that included games on the July 1, 2008, list were computed for all taking the survey up to June 30. For participants taking the survey from July 1 to September 30, games counts up to the October 1 list were used, and for those taking the survey after October 1, games counts up to and including the January 1, 2009, list were used. About 50% of the participants taking the survey after June 30 had games counts of zero in the October 1 and January 1 lists, so there was little systematic effect of the slight differences.

Many players played relatively few games, and their recall task may have been easy. It might be quite easy to recall that one has played only 7 games if one has played in only one international tournament and the tournament was this year. So, a supplementary analysis was done for the 103 participants who had played 200 games or more. The 200-games value was selected to give a sizeable games count.

The accuracy of two methods of estimating total games played was compared. Participants made an overall total estimate of games and an estimate of number in a typical year. The author then multiplied the latter value by time from domain entry to the time of taking the survey, using a participant's estimated entry year.

Results

Entry year estimates

Table 1 presents raw accuracy data for the year of entering the domain. The distributions of actual and estimated years in the domain depart from normality, defined as kurtosis and skewness values between 1 and −1 (For actual year, kurtosis = 2.12 and skewness = −1.44, and for estimated year, kurtosis = 1.87, skewness = −1.4). For further analysis, each value for actual and estimated entry year was converted to a value for years in the domain (e.g., if the actual entry year was 2006, the years-in-the-domain value was 2008 – 2006 = 2). A value of 1 was added to each years-in-the-domain value, to avoid log-transforming 0 s, and then the common logarithm was calculated; for instance, $\log(X + 1)$, or specifically, $\log(2008 - 2006 + 1) = 0.48$. Figure 1 gives a scatterplot of the log-transformed actual and estimated years in the domain.

About 57.34% of all 668 participants correctly named their entry year, and about 36.64% of participants first on the list before 2001 did so. The median signed accuracy value is 0, while the mean value shows an estimated entry date 0.55 years later than actual entry. For participants first on the list before 2001, the median is 0 and the mean signed estimated date is nearly 1 year later than the actual date. Examining the absolute values for all players, the median is 0 but the mean is nearly 1. For players entering the list before 2001, the median is 1 and the mean about 1.5. Table 2 presents correlations between log-estimated and actual values for the year and games estimates. Table 2 shows that the correlation between transformed log values of the actual and recalled years in the domain for all participants was high, at .97 ($p < .001$), and for those first on the list before 2001 it was .92 ($p < .001$).

Figure 2 presents the mean and median signed and absolute raw accuracy values in the year categories. The figure suggests a trend toward declining accuracy until the 25+ years category. The signed and absolute values for the accuracy measures departed greatly from normality. For the absolute value, kurtosis was 16.37 and skewness was 3.86. The concern here was with absolute values, and these were transformed to give an approximate normal distribution. A value of 1 was added to the absolute value of each accuracy measure, and then the common log was taken—that is, $\log(X + 1)$. The transformed distribution was a more acceptable approximation to normality (kurtosis = 1.04, skewness = 1.27).

Figure 3 presents log-transformed descriptive statistics for the absolute values of the accuracy measures. Figure 3 suggests declining mean and median accuracy values until

Table 1 Accuracy of year estimates for entering the domain, where accuracy = actual year – recalled year (e.g., 2002 – 2004 = −2)

	All participants (n = 668)		Participants in before 2001 (n = 262)	
	Absolute value	Signed value	Absolute value	Signed value
Median	0	0	1	0
Mean	0.92	–0.55	1.53	–0.89
SD	1.72	1.87	2.12	2.47
SE	0.07	0.07	0.13	0.15
Range	0 to 15	–15 to 10	0 to 12	–12 to 10

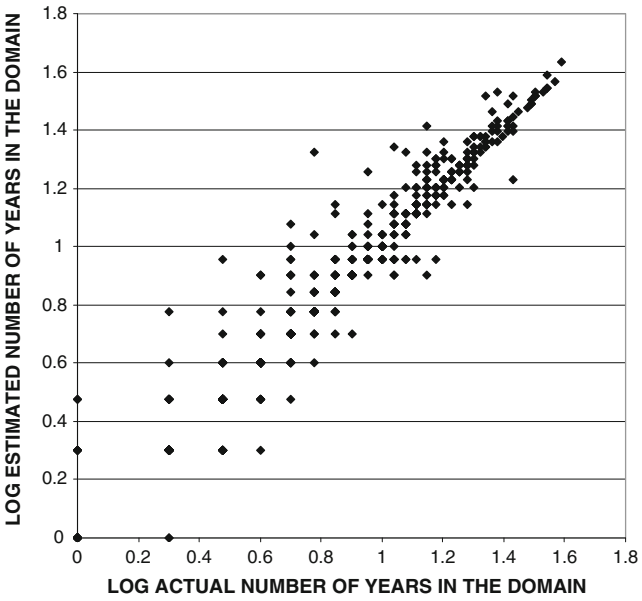

Fig. 1 Scatterplot of log estimated and actual numbers of years in the domain for all participants

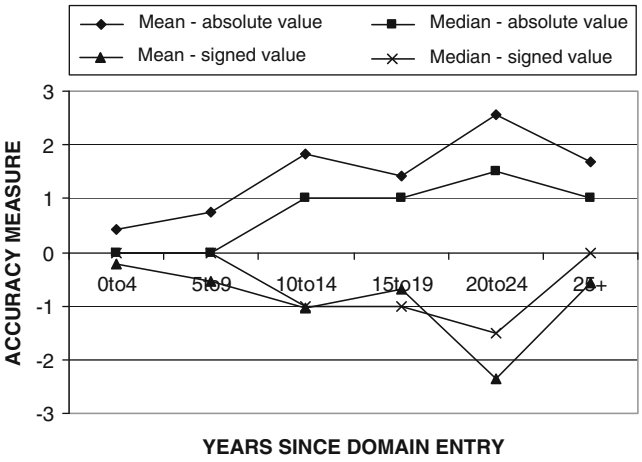

Fig. 2 Mean and median raw accuracy scores for the years-in-domain categories. A negative value signifies a recalled entry year later than the actual entry year

the 10–14 years category, and then no systematic further decline in accuracy. The effect of category was significant [$F(5, 662) = 27.89$, $p < .001$, adjusted $R^2 = .17$]. Post hoc Tukey HSD tests found no significant differences between estimates in the 0–4 and 5–9 categories ($SE = 0.022$, $p = .102$) but significant differences between those in the 0–4 category and the remaining categories. Of these differences, the smallest mean difference was that between 0–4 and 15–19, which was significant ($SE = 0.034$, $p < .001$). The largest difference in the four categories from 10–14 games and larger was that between 15–19 and 20–24, which was not significant ($SE = 0.061$, $p = .32$). Figure 3 suggests increasing *SD*s and *SE*s over time.

Games estimates

The distributions for the actual and estimated FIDE-rated game total and games per year values departed from normality, with kurtosis and skewness measures all outside the range −2 to 2. All were log-transformed with 1 added to each raw score to eliminate zeros—that is, the common log of $(X + 1)$ was taken; for instance, specifically, log(0 total games + 1) = 0. All transformed distributions approximated normality, with kurtosis and skewness measures between 1 and −1.

Games in a typical year

Table 3 presents games-per-year data for all participants, with actual values calculated from the database and estimated values from the Question 4 data. The percent difference between actual and recalled values was calculated as follows: (actual value – recalled value) × 100/actual value.

Figure 4 presents a scatterplot of the log-transformed estimated and actual games-per-year values for all participants. For all players, the correlation between the log-transformed actual and estimated games per year was .66 ($p < .001$), and for the 200 games group it was .63 ($p < .001$).

For all participants, from the totals, the median difference between actual and estimated games per year was 29.14%. From individual scores, the median absolute-value difference was 38.89%. However, 200-games group participants on average showed very little difference between actual and estimated values.

Table 2 Correlations between estimated and actual values

	Years in domain	Games per year	Total games	Total games by average year
All players ($n = 668$)	.97*			
Group entering before 2001 ($n = 262$)	.92*			
All players ($n = 629$)		.66*	.88*	.78*
Group which played 200 games ($n = 103$)		.63*	.66*	.60*

* $p < .001$

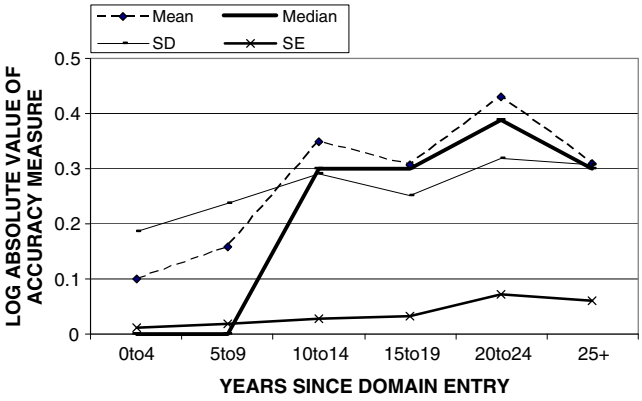

Fig. 3 Means, medians, *SD*s, and *SE*s for log absolute values of accuracy scores in the different year categories

Total games estimates

From the total-games counts (Question 5), about 5.56% of the 629 participants got the exact games count correct. None in the 200-games group did. For all players, the correlation between the log-transformed values was .88 ($p < .001$), and for the 200-games group it was .66 ($p < .001$). Tables 4 and 5 give participants' actual total-games values and estimates from the Question 5 data. Figure 5 gives a scatterplot. For all participants, the percentage difference from median totals was 16.67%, and from individual scores the absolute value of the difference was 25.49%. However, the 200-games group had a median 11.82% difference from the actual totals, and from the absolute values of individual scores there was a 25.93% difference.

Total-games estimates from games in a typical year and years in the domain

The lower parts of Tables 4 and 5 present estimated total games calculated by multiplying participants' estimated games per year by estimated years in the domain (derived from participants' estimated year of appearing on the rating list). So, if a participant took the survey in 2008, estimated their list entry year as 2000, and estimated playing 10 games in a typical year, the total-games count was (2008 – 2000) × 10, or 80 games. A constant of 1 was added to all final games totals to eliminate zeros.

For all players, the correlation between the log-transformed actual and estimated totals was .78 ($p < .001$), and for the 200-games group it was .6 ($p < .001$). For all players, the median value from totals was a difference from the actual games totals of 15%, and from individual scores the median difference was 41.18%. For the 200-games group, from totals the median difference was 11.82%, and from individual scores it was 31.64%.

Which method of estimating total games played was more accurate? From the medians of the absolute values of individual scores for all players and the 200-games group, the total-games estimates from Question 5 were more accurate.

General inaccuracy?

Are some participants generally less accurate? Were participants who were inaccurate in their entry-year estimates also inaccurate in the total-games and games-per-year estimates? The data of the 629 participants with

Table 3 Actual and estimated games per year

	Actual	Recalled	Percent difference		
			From totals (Ab. value)	From individual scores	
				Absolute value	Signed value
All Participants (n = 629)					
Median	15.52	11	29.14	38.89	11.93
Mean	20.93	17.47	16.48	61.97	–10.80
SD	19.75	18.67		133.89	147.16
SE	0.79	0.74		5.34	5.87
Range	0 to 168	0 to 175		0 to 1791.67	–1791.67 to 100
200-Games Group (n = 103)					
Median	29.31	30	2.36	4.63	0.38
Mean	35.69	40.28	12.87	8.73	0.14
SD	17.57	29.16		9.17	12.69
SE	1.73	2.87		0.9	1.25
Range	11.29 to 77.81	1 to 175		0 to 37.82	–37.82 to 30.83

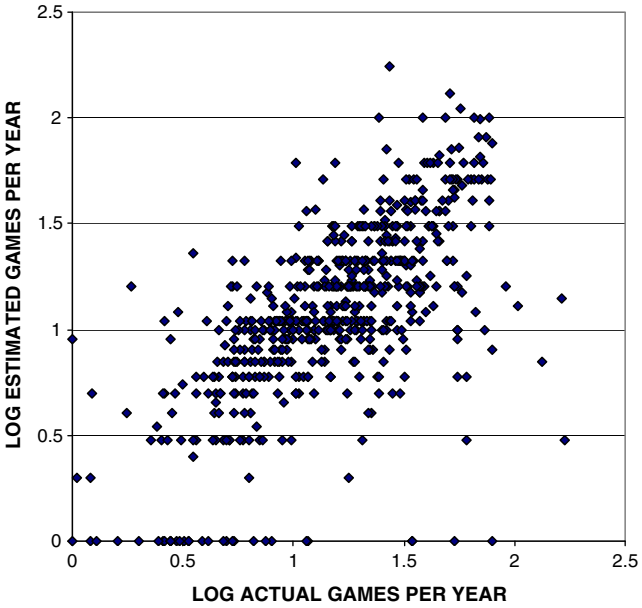

Fig. 4 Scatterplot of log estimated and actual numbers of games per year for all participants

estimates for entry year, total games, and games per year were examined. For each participant, common logarithm values (plus 1) of the total-games estimates (from Question 5) and games-per-year estimates (plus 1) were calculated. Similar log values (plus 1) were calculated for the actual values. For each measure, the estimated log value was subtracted from the actual log value, and the absolute value of the difference was calculated. This calculation also was done for years in the domain; that is, the log of estimated years plus 1 was subtracted from the log of actual years plus 1, and the absolute value was calculated. The three difference values for each participant were then correlated. There was some evidence that some participants were less accurate in general, with a low positive correlation between inaccuracy in year estimates and in total-games counts of .13 ($p < .001$) and a correlation between inaccuracy in total-games counts and in games in a typical year of .24 ($p < .001$). However, the correlation between inaccuracy in year estimates and in games in a typical year of .02 was not significant ($p = .61$).

Discussion

The key findings may be summarized as follows. On average, participants were reasonably accurate at estimating entry year, and expertise researchers can be more confident about other career date estimates. Estimates of games in a typical year were also reasonably accurate. The correlations between the actual and estimated values for entry year, games in a typical year, and total games were often higher than the correlations of .6 to .74 or more that have been found in some diary studies between the diary and past-year estimates of the extent of practice. Asking for an overall total number of games was more accurate than asking for an estimate for a typical year and multiplying that by a participant's estimated years in the domain. However, some participants were quite inaccurate.

Limitations of the present study

The present study has limitations and may be criticized on various grounds. First, one might query the focus on absolute values. As noted, these are preferred in accuracy

Table 4 Games-count estimates of all participants ($n = 629$)

	Actual	Recalled	Percent difference		
			From totals (Ab. value)	From individual scores	
				Absolute value	Signed value
Total-Games Estimate					
Median	60	50	16.67	25.49	0
Mean	116.56	125.55	7.71	53.87	–28.19
SD	157.1	195.46		155.41	162.06
SE	6.24	7.8		6.2	6.46
Range	1 to 1140	1 to 2000		0 to 1900	–1900 to 92.86
Total Games From Games-in-a-Typical-Year Estimates					
Median	60	51	15	41.18	7.69
Mean	116.56	135.72	16.44	69.5	–19.12
SD	157.11	255.26		158.66	172.17
SE	6.24	10.18		6.32	6.86
Range	1 to 1140	0 to 2975		0 to 1929.41	–1929.41 to 100

Table 5 Games-count data of participants who had played at least 200 games (n = 103)

	Actual	Recalled	Percent difference		
			From totals (Ab. value)	From individual scores	
				Absolute value	Signed value
Total-Games Estimate					
Median	313	350	11.82	25.93	–3.81
Mean	396.31	427.62	7.90	34.23	–8.34
SD	211.61	320.29		31.31	45.75
SE	20.85	31.56		3.08	4.51
Range	200 to 1140	20 to 2000		0.33 to 150	–150 to 92
Total Games From Games-in-a-Typical-Year Estimates					
Median	313	350	11.82	31.64	–0.48
Mean	396.31	475.97	20.10	46.81	–17.84
SD	210.83	469.01		73.47	85.38
SE	20.85	46.21		7.24	8.41
Range	200 to 1140	14 to 2975		0.2 to 532.98	–532.98 to 94.4

research, because signed values may partially cancel each other out. Second, one might query the use of separate groups (e.g., participants on the list before 2001 and those having played at least 200 games). Many participants entered the list very recently or had very low games counts, and their recall task was very easy. Data for all participants are presented, and any researcher who believes that the use of a supplementary group is unsound should focus just on the data for all participants.

Third, as mentioned above, the FIDE data are not perfect, but with the large sample size any impact of this would be negligible. Players deliver a game result to the tournament director, who forwards the game results to the national federation, which forwards it to FIDE. Inaccuracy in FIDE data is not linked to memory inaccuracy in players.

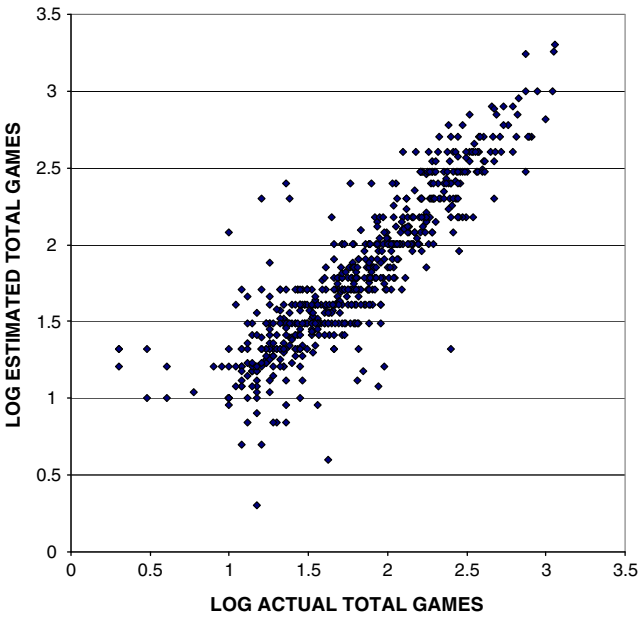

Fig. 5 Scatterplot of the log estimated and actual game totals for all participants

Fourth, one might argue that participants may have just looked up their entry date and games counts, despite the instruction in two questions to answer from memory only. It is unlikely that anyone did this in the middle of a 23-question online survey, whose main focus was on practice and natural talent. Furthermore, the participants who entered the list before 2001 and/or who had played 200 games or more had no readily available way to look up those data. For the others, FIDE publishes lists from July 2001 on its website, but retrieving an entry date would require downloading and trawling through many lists. Similarly, trying to get games counts would meet difficulties. In addition, the data themselves suggest that the participants did not look up the data. As mentioned, many participants were in the right ballpark for accuracy, but those in the 200-games group and those first on the list before 2001 mostly did not have the exact values.

Fifth, as mentioned, expertise researchers have not been much concerned with such career data. But, a reasonable hypothesis is that estimating FIDE domain entry date is quite comparable to estimating dates of learning the chess moves and of the start of serious practice. These dates generally are not reported on any forms (e.g., for joining chess clubs) and are not conversation topics among chess players. Most participants were reasonably adept at estimating year of list entry, consistent with findings that people are reasonably adept at estimating dates of historical events. But the games estimates were more problematic. How far one can extrapolate accuracy of games estimates to accuracy of estimates of amount of practice? Players were

reasonably accurate at estimating the number of games in a typical year, and probably are reasonably accurate at estimating the number of practice hours in a typical year. But as mentioned, the true values for practice hours for international chess players mostly are unknowable, and estimates never can be checked against reality. Games counts are known, and estimates can be cross-validated with a measure whose true values are known. Wildly inaccurate games estimates here would cast doubt on the accuracy of analogous practice estimates.

Implications of the present study

The present data have methodological implications for expertise researchers. The first is to focus more on the date estimates of some participants. Perhaps some just have poor memory for career data. Indeed, there was some evidence that those inaccurate at year estimates were also inaccurate at total-games estimates. Individuals' abilities to form and retrieve episodic memories vary widely (Kirchhoff, 2009; Malmberg, 2007), and accuracy may be affected by depth of processing and emotional context during encoding, as well as by such factors as field independence (Corson, Verrier, & Bucic, 2009). Other explanations are that some participants may have misunderstood the questions or made incorrect inferences. The participants knew that they had a rating because it was a requirement for study participation. But, because FIDE does not directly inform players when they enter the list, some may simply not have known and may have inferred it. Even if they had not been informed directly by FIDE, a player would know that he or she had a rating when participating in the next international tournament after being rated: Tournament directors retrieve published ratings and rank all players in the event by those ratings for seeding purposes. Unrated players would be so designated on the tournament pairing table, scrutinized by players every round.

A second methodological implication is that expertise researchers should not just present mean and signed values. Mean values can be greatly inflated by the extreme scores of a few participants. The author suggests using, or at least reporting, median values to deal with outliers.

A third implication is that the method of estimating practice totals from an estimate in a typical year may yield sound data, but it probably is better to ask participants to make an overall estimate. Further research might examine the accuracy of year-by-year estimates and the value of extrapolating a past-year estimate.

How might career data estimates be improved in expertise studies? Much research has examined various sources of unreliability in answering survey questions and various ways to improve them (Beatty & Willis, 2007; Collins, 2003). Respondents may misunderstand questions or may be unsure of some word meanings, and thus may provide answers that only appear legitimate (Collins, 2003). Cognitive interviews can be used to pretest and revise questions in order to reduce misinterpretations, and adding an "I do not know" option may also increase reliability.

In the present study, one source of unreliability may have been incorrect inferences. An example is a participant on the inaugural 1970 rating list who gave his year of first rating as 1966, 4 years earlier. Tournament records showed that he indeed played in his first international tournament in 1966, and he evidently inferred (incorrectly) that he gained a rating immediately afterward. Also, some players may have confused FIDE-rated with nonrated games played in local tournaments. Participants also may use different recall strategies, with varying accuracy. For example, if asked in which year they entered the FIDE list, a player might try to recall mental landmarks (was I in school or working at my first job?), or might try to determine the age at which it occurred and then calculate the year. The strategies used in retrospective recall have been little investigated by expertise researchers. Further research might investigate what strategies are used and the possible impacts of varying strategies on accuracy, perhaps by using the FIDE database or diary data. Do particular strategies consistently yield more accurate data? Participants in expertise studies then might be asked to use such strategies in recall. For instance, De Bruin et al. (2008) asked participants to try to recall data year by year by giving cues. Reimer and Matthes (2007) also suggested some ways to improve recall of autobiographical data.

Expertise researchers also might try to identify and exclude very inaccurate participants. One way to do so would be to include some prior episodic memory test. A study with international chess players, for instance, might check the accuracy of their domain entry and games-count estimates. Or, details of previous games and opponents, as in the above Campitelli et al. (2008) study, might be used. Researchers in autobiographical memory also might use archival databases as a useful research method (Campitelli et al., 2008).

In conclusion, the present data show that most participants are reasonably accurate in recall of career data, giving some confidence in the accuracy of other career data from the retrospective recall paradigm. But some were quite inaccurate, and researchers might try to identify and exclude such participants.

References

Bahrick, H. P. (1984). Semantic memory content in permastore: Fifty years of memory for Spanish learned in school. *Journal of Experimental Psychology. General, 113,* 1–24.

Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly, 71,* 287–311.

Birnbaum, M. H. (2004). Human research and data collection via the Internet. *Annual Review of Psychology, 55,* 803–832.

Brown, N. R., Rips, L. J., & Shevell, S. K. (1985). The subjective dates of natural events in very long term memory. *Cognitive Psychology, 17,* 139–177.

Burt, C. D. B., Kemp, S., & Conway, M. (2004). Memory for true and false autobiographical memory descriptions. *Memory, 12,* 545–552.

Campitelli, G., Parker, A., Head, K., & Gobet, F. (2008). Left lateralization in autobiographical memory: An fMRI study using the expert archival paradigm. *International Journal of Neuroscience, 118,* 191–208.

Charness, N., Tuffiash, M., Krampe, R., Reingold, E., & Vasyukova, E. (2005). The role of deliberate practice in chess. *Applied Cognitive Psychology, 19,* 151–165.

Christianson, S. (1989). Flashbulb memories: Special, but not so special. *Memory & Cognition, 17,* 435–443.

Collins, D. (2003). Pre-testing survey instruments: An overview of cognitive methods. *Quality of Life Research, 12,* 229–238.

Colvin, G. (2008). *Talent is overrated*. New York: Portfolio.

Corson, Y., Verrier, N., & Bucic, A. (2009). False memories and individual variations: The role of field dependence-independence. *Personality and Individual Differences, 47,* 8–11.

Cumming, J., Hall, C., & Starkes, J. L. (2005). Deliberate imagery practice: The reliability of using a retrospective recall methodology. *Research Quarterly for Exercise and Sport, 76,* 306–323.

Davids, K. (2000). Skill acquisition and the theory of deliberate practice: It ain't what you do it's the way that you do it! *International Journal of Sport Psychology, 31,* 461–466.

De Bruin, A. B. H., Smits, N., Rikers, R. M. J. P., & Schmidt, H. G. (2008). Deliberate practice predicts performance over time in adolescent chess players and drop-outs: A linear mixed models analysis. *British Journal of Psychology, 99,* 473–497.

Deakin, J. M., & Cobley, S. (2003). An examination of the practice environments in figure skating and volleyball: A search for deliberate practice. In J. Starkes & K. A. Ericsson (Eds.), *Expert performance in sports: Advances in research on sport expertise* (pp. 90–113). Champaign: Human Kinetics.

Deakin, J. M., Cote, J., & Harvey, A. S. (2006). Time, budgets, diaries and analyses of concurrent practice activities. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 303–318). New York: Cambridge University Press.

Elo, A. E. (1986). *The rating of chess players, past and present* (2nd ed.). New York: Arco.

Ericsson, K. A. (2006). The influence of experience and deliberate practice on the development of superior expert performance. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 683–703). New York: Cambridge University Press.

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

Friedman, W. J. (1993). Memory for the time of past events. *Psychological Bulletin, 113,* 44–66.

Gladwell, M. (2008). *Outliers*. Boston: Little, Brown.

Gobet, F., & Campitelli, G. (2007). The role of domain-specific practice, handedness and starting age in chess. *Developmental Psychology, 43,* 159–172.

Hodges, N. J., & Starkes, J. L. (1996). Wrestling with the nature of expertise: A sport specific test of Ericsson, Krampe and Tesch-Romer's (1993) theory of deliberate practice. *International Journal of Sport Psychology, 27,* 400–424.

Howard, R. W. (2006). A complete database of international chess players and performance ratings for varied longitudinal studies. *Behavior Research Methods, 38,* 698–703.

Howard, R. W. (2009). Individual differences in expertise development over decades in a complex intellectual domain. *Memory & Cognition, 37,* 194–209.

Kemp, S. (1988). Dating recent and historical events. *Applied Cognitive Psychology, 2,* 181–188.

Kirchhoff, B. A. (2009). Individual differences in episodic memory: The role of self-initiated encoding strategies. *The Neuroscientist, 15,* 166–179.

Larsen, S. T., & Conway, M. A. (1997). Reconstructing dates of true and false autobiographical memories. *European Journal of Cognitive Psychology, 9,* 259–272.

Malmberg, K. J. (2007). Toward an understanding of individual differences in episodic memory: Modeling the dynamics of recognition memory. In A. S. Benjamin & B. H. Ross (Eds.), *The psychology of learning and motivation: Skill and strategy in memory use* (pp. 313–349). San Diego: Academic Press.

Milavsky, J. R. (1992). How good is the A. C. Nielsen people-meter system? *Public Opinion Quarterly, 56,* 102–115.

Niedzwienska, A. (2003). Distortion of autobiographical memories. *Applied Cognitive Psychology, 17,* 81–91.

Potts, R., Belden, A., & Reese, C. (2008). Young adults' retrospective reports of childhood television viewing. *Communication Research, 35,* 39–60.

Reimer, M., & Matthes, B. (2007). Collecting event histories with TrueTales: Techniques to improve autobiographical recall problems in standardized interviews. *Quality and Quantity, 41,* 711–735.

Ross, P. E. (2006). The expert mind. *Scientific American, 295,* 64–71.

Starkey, G. (2004). Estimating audiences: Sampling in television and radio audience research. *Cultural Trends, 13,* 3–25.

Tekcan, A. I., Ece, B., Gulgoz, S., & Er, N. (2003). Autobiographical and event memory for 9/11: Changes across one year. *Applied Cognitive Psychology, 17,* 1057–1066.

Williams, H. L., Conway, M. A., & Cohen, G. (2008). Autobiographical memory. In G. Cohen & M. A. Conway (Eds.), *Memory in the real world* (pp. 21–90). New York: Psychology Press.

Wright, D. B. (1993). Recall of the Hillsborough disaster over time: Systematic biases of "flashbulb" memories. *Applied Cognitive Psychology, 7,* 129–138.