A psychometric analysis of chess expertise

HAN L. J. VAN DER MAAS AND ERIC-JAN WAGENMAKERS
University of Amsterdam

This study introduces the Amsterdam Chess Test (ACT). The ACT measures chess playing proficiency through 5 tasks: a choose-a-move task (comprising two parallel tests), a motivation questionnaire, a predict-a-move task, a verbal knowledge questionnaire, and a recall task. The validity of these tasks was established using external criteria based on the Elo chess rating system. Results from a representative sample of active chess players showed that the ACT is a very reliable test for chess expertise and that ACT has high predictive validity. Several hypotheses about the relationships between chess expertise, chess knowledge, motivation, and memory were tested. Incorporating response latencies in test scores is shown to lead to an increase in criterion validity, particularly for easy items.

There are two main reasons to study the psychology of chess (de Groot & Gobet, 1996). First, the skill of top chess players is an excellent example of cognitive expertise. When people practice a couple of hours per day for many years they can reach amazing levels of expertise in sports, science, or arts (Charness, 1991). Many facts about expertise were first discovered in the domain of chess (Ericsson, 2003). Second, chess often is studied as an example of higher-order cognition. Several researchers, most notably Newell and Simon (1972), used chess playing as an ecologically valid yet controlled environment for the study of higher-level reasoning.

An advantage of using chess as an environment in which to study expertise or higher-level cognition is that the level of expertise can be rigorously quantified by a rating system based on Thurstone's case V model (Batchelder & Bershad, 1979; Thurstone, 1994). Elo (1978) applied this rating system to chess playing, and today almost all club players have an "Elo rating" to quantify their playing strength. Figure 1 shows a frequency distribution of Elo ratings for all members of the Royal Dutch Chess Federation (KNSB). The distribution is approximately normal. Three or four very good players have a rating higher than 2,600, and the weakest players have a rating below 800. The outcome of a game between two players can be predicted, with varying degrees of accuracy, from the difference between their prior Elo ratings. For example, if A plays B, and A has 400 rating points more than B, the expected score of A is .92 (includ-

AMERICAN JOURNAL OF PSYCHOLOGY
Spring 2005, Vol. 118, No. 1, pp. 29–60

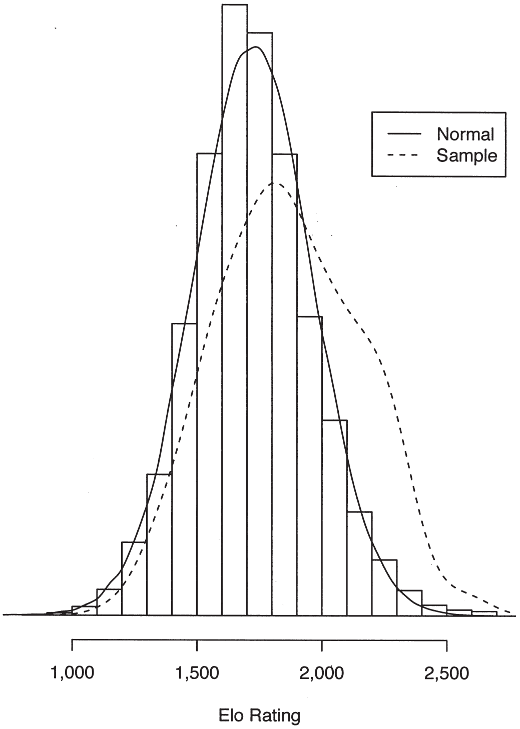


Figure 1. Frequency distribution of Elo ratings for the members of the Dutch chess federation (*N* = 9,109, May 1998) compared with normal distribution and the sample used for the ACT

ing draws; wins, draws, and losses are scored 1, .5, and 0, respectively). It takes about 25 games (Dutch rating system), each possibly lasting up to 7 hr, to obtain a reliable Elo rating, and ratings are updated four times a year. For the Dutch rating system, Jonker (1992) estimated the standard error of a grand master's Elo rating at 25 and the standard error of an average player's Elo rating at 53.

Chess research on both expertise and cognition has been strongly influenced by the pioneering work of de Groot (1946/1978). De Groot collected thinking-aloud protocols of top chess players and analyzed their thought processes. One of de Groot's results concerned the phenomenal memory of top chess players for chess positions and motivated an influential line of research starting with the studies of Simon and Chase (1973). Although the theory of Simon and Chase has come under attack for several reasons (for reviews, see Charness, 1991; de Groot & Gobet, 1996, chapter 5; Ericsson & Lehmann, 1996; Saariluoma, 1995), it nonetheless encouraged the development of a number of alternative theories and research methods.

Over the last 15 years psychological research on chess playing has gained momentum.[1] Research in chess now uses methods such as eye tracking (de Groot & Gobet, 1996; Charness, Reingold, Pomplun, & Stampe, 2001), positron emission tomography (Nichelli et al., 1994), magnetoencephalography (Amidzic, Riehle, Fehr, Wienbruch, & Elbert, 2001), and functional magnetic resonance imaging (Atherton, Zhuang, Bart, Hu, & He, 2003). Howard (1999) and Gobet, Campitelli, and Waters (2002) discussed the relationship to IQ, especially regarding the Flynn effect. In addition, theories that account for decision-making processes of expert chess players are continuously developed and adjusted (de Groot & Gobet, 1996; Gobet, 1997; Ericsson & Kintsch, 1995; Ericsson, Patel, & Kintsch, 2000; Saariluoma, 1995). Reviews by Ericsson and Smith (1991), Charness (1991), and Ericsson and Lehmann (1996) showed how numerous factors and processes involved in chess playing have been successfully identified. However, partly for methodological reasons, several issues are still unresolved (Chabris & Hearst, 2003). In most chess studies, sample sizes are very small, and selection of participants and test materials is not standardized (cf. Gobet, De Voogt, & Retschitzki, 2004). We believe that a standardized psychometric analysis of chess playing ability, of which the current study is the first step, is a fruitful new approach.

The development of a psychometric test for chess proficiency, such as the Amsterdam Chess Test (ACT) presented here, was motivated by several specific questions. A first question is whether it is possible to measure chess ability by using the simple test formats used in standard ability tests. In the study of expertise, it is vital to use representative tasks that allow replication in a laboratory setting (Ericsson & Smith, 1991; Ericsson, 2003). The extent to which psychometric tests are able to capture playing strength outside the context of a competitive, on-the-board playing situation is largely unknown. To our knowledge, no comprehensive psychometric tests for chess proficiency exist (a possible exception is Pfau & Murphy, 1988). Validation of such a test of chess-playing expertise can be rigorously accomplished by using the Elo ratings of the participants as an external performance criterion. In contrast, many other psychological tests (e.g., psychometric IQ tests) lack such a well-defined criterion.

A second question concerns the degree to which certain test formats and psychometric measures are more useful than others in measuring chess expertise. The availability of a strong external criterion allows a set of different test formats and psychometric methods (e.g., item response models, adaptive testing, methods of differential item functioning) to be analyzed with respect to their predictive validity.

In addition, we believe that a psychometric analysis of chess expertise can provide useful information not only about chess expertise per se but also more generally about the adequacy of certain psychometric techniques. The presence of a strong external criterion and the fact that chess playing is a highly constrained yet ecologically valid cognitive skill make the study of chess playing an excellent testing ground for the evaluation of psychometric techniques. Thus, chess playing might function as a *Drosophila* not only for cognitive psychology (Simon & Chase, 1973) but also for psychometrics.

A third question is whether a psychometric ability test for chess can provide information about the kind of processes that underlie chess proficiency (Ericsson & Smith, 1991). The ACT was constructed to address several theoretical questions about the nature of chess proficiency. For example, a distinction is commonly made between three chess skills. Tactical ability entails the discovery and accurate calculation of combinations. Usually a combination means that material (i.e., a pawn or a piece) is temporarily sacrificed to achieve certain greater short-term goals. In contrast to tactical ability, positional insight or judgment entails little calculation of concrete sequences of moves and countermoves. Instead, the focus is on strengthening one's position through moves that are based on general principles and result in long-term payoffs. Finally, endgame knowledge refers to procedures to handle standard endgame situations. The distinction between these three different factors in chess ability bears some resemblance to the debate about "*g*" in intelligence research, that is, the discussion whether intelligence is based on some general ability or on more specific abilities.

Finally, the existence of a reliable and valid test for chess-playing strength has many practical applications. For instance, in several situations (e.g., in tournaments or in the thriving Internet chess community) it is difficult to pair a player to a suitable opponent without a valid rating of the player. For chess training, which in Holland may be part of the school curriculum, it is sometimes desirable to assign players to different groups or to different training programs according to skill. Many younger players do not have a reliable Elo rating, and a fast and reliable estimation of playing strength could prove very useful in these circumstances. Moreover, the effects of training could also be evaluated with help of a reliable and valid chess test.

The ACT

The ACT is a computerized task that consists of several subtests. The choice of subtests was motivated by a number of arguments. The main subtests, choose-a-move A and B, are two parallel tests that consist of chess problems commonly published in chess magazines and books. These chess problems require the participant to find a single best move from a chess diagram that depicts a particular position. Choose-a-move tasks have a long tradition in chess psychology (de Groot, 1946/1978). Based on the practice of standard ability testing in psychology, we expected that a test consisting of a series of such problems would yield a very good test of chess proficiency (see also Ericsson & Kintsch, 1995, p. 233). However, several top players and trainers questioned our expectation, so we included other tests in the ACT. These other tests are motivated partly by previous empirical and theoretical work in the psychology of chess playing and partly by comments of expert chess players and chess trainers.

Many chess experts believed that the common tactical chess problem for which the key move leads to a fast and spectacular win would differentiate well only in the lower Elo rating range (i.e., for mediocre players). The differences between players in the higher Elo rating ranges were believed to be more related to positional and endgame knowledge. To test this hypothesis, the choose-a-move tests incorporated tactical, positional, and endgame problems.

A potential drawback of the choose-a-move test may be a lack of ecological validity. The items in the choose-a-move test were exceptional in that a single move (i.e., the solution) was much better than all other legal moves, whereas this is rarely the case in a practical chess game. As an alternative training exercise, chess magazines sometimes publish tests in which readers have to predict the sequence of moves that was played in a single game. Points can be scored not only by correct prediction of the move actually played in the game but also by the suggestion of a good alternative move. Because of its ecological validity and its popularity, a predict-a-move test was added to the ACT.

The importance of the right motivation (i.e., all favorable psychological factors not directly related to technical knowledge of the game such as competitive spirit and the ability to remain calm under pressure) is widely acknowledged throughout the chess world and in professional sports. Especially in matches between top players, psychological factors are thought to exert a substantial effect on the outcome. One of the most famous chess trainers today, Mark Dvoretsky, explicitly argues that success in chess is the product of ability and motivation (Dvoretsky, personal communication, October 2002). In this context, de Groot (1946/1978) mentions the work of Djakow, Petrovsky, and Rudik (1926), who used the Rorschach test to demonstrate that grand masters have a high "will power." More in line with current fashion, Joireman, Fick, and Anderson (2002) showed that chess players score higher on tests of sensation seeking than do controls. To determine the importance of motivation, a chess motivation questionnaire was added to the ACT.

Chess psychology historically is strongly oriented toward the study of memory and recall of chess positions. In the current literature, models of chess memory are the topic of a lively debate (Chase & Simon, 1973; Ericsson & Kintsch, 1995; Gobet, 1998, 2000; Robbins et al., 1996; Saariluoma & Kalakoski, 1998; Vicente & Wang, 1998). For instance, strong chess players can play several games simultaneously without seeing the actual positions (i.e., blindfold play). Playing one complete game from the mind's eye (without committing gross errors) already places a burden on human memory because the necessary calculation of future moves can interfere with the representation of the actual position. Obviously, the addition of other games increases memory load even further in that representations for different games can also interfere with each other. Ericsson and Kintsch (1995) argued that the standard short-term and long-term memory models cannot easily explain the phenomenon of simultaneous blindfold play.

More than a century ago the phenomenon of blindfold play also stood at the basis of the psychological study of chess (Binet, 1893/1966). The monograph of Binet, who was in many ways the founder of the psychometric approach that is pursued in this article, was a major inspiration for de Groot (1946/1978) to perform his famous chess studies starting in 1938. De Groot concluded that individual differences in chess expertise did not result from a differential capacity for calculating many moves ahead (e.g., depth of search). Using spoken (i.e., think-aloud) protocols during choose-a-move tasks, de Groot argued that skill differences are instead associated with performance in recall and recognition of standard chess positions. Skilled players presumably calculate as many moves as unskilled players, but the recognition of familiar chess patterns that drives the move selection process allows skilled players to exclude bad moves and focus their efforts on promising continuations. In his later work, de Groot described this skill difference in terms of chess intuition (de Groot, 1986).

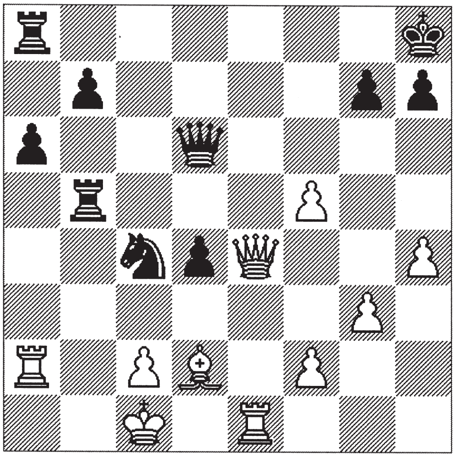
Although the recall and recognition performance of chess experts is indeed amazing, the conceptualization of skill differences in terms of performance on memory tasks has been qualified (Holding & Pfau, 1985). Subsequent research demonstrated a skill effect on depth of search (see Charness, 1991, and Gobet et al., 2004, for reviews). Furthermore, Charness (1981a, 1981b) criticized de Groot's results on the basis of the small and nonrepresentative sample used. A famous result that plays a major role in this discussion is the absence of an expertise effect in recall for random positions (Simon & Chase, 1973; see also Vicente & de Groot, 1990). However, this result has also been disputed. Saariluoma (1989), Gobet and Simon (1996), and Gobet and Waters (2003) demonstrated that a small skill effect is still found for random positions. To test how well memory of chess positions explains the differences in chess ability (Ericsson et al., 2000) and to replicate the attenuation of the skill effect caused by randomness of chess position, a chess recall test was added to the ACT.

Finally, Pfau and Murphy (1988) claimed that verbal chess knowledge is an important determinant of chess ability rather than a byproduct of it. Pfau and Murphy assessed the predictive value of verbal chess knowledge in a large group of chess players (*N* = 59). Although verbal chess knowledge correlated significantly with Elo rating (*r* = .69), Pfau and Murphy could not demonstrate an independent predictive value of verbal knowledge when choose-a-move tests were incorporated in the regression equation. Nonetheless, Pfau and Murphy claimed that verbal knowledge is a determinant and not an incidental correlate of chess skill (see also Holding, 1985). To investigate this hypothesis we added a verbal knowledge questionnaire (cf. Pfau & Murphy, 1988) to the ACT.

In sum, the choose-a-move tests A and B form the core of the ACT. For reasons just outlined the ACT was supplemented with four other tests: a predict-a-move test, a chess motivation questionnaire, a recall test for chess positions, and a verbal knowledge questionnaire. The ACT subtests are discussed in more detail in the sections that follow. The details of the origin and scoring of all ACT subtests items are available at http://users.fmg.uva.nl/hvandermaas/chesshtml/act.htm. The ACT was improved by a short period of pilot testing that featured about 15 chess players of various skill levels.

Choose-a-move test A and B

The choose-a-move test B was designed to be a parallel version of the choose-a-move test A. Test A and test B both consisted of three sets of items. The first set consisted of 20 tactical items, the second set consisted of 10 positional items, and the third set consisted of 10 endgame items. Item difficulty increased within each set. Three easy items were used to familiarize the participants with the choose-a-move test format. Most items were chess problems taken from Bloch (1994), Suetin (1976), and Portisch and Srközy (1986). Several items were adapted or newly constructed by the second author. An important selection criterion was that one move was clearly superior to all other moves. The left part of Figure 2 shows an example item, a graphic representation of a chess position (i.e., a chess diagram). Items were presented on a computer screen for 30 s or until the participant responded. While the participant was solving an item, a small clock located below the diagram displayed the time that remained before the allotted 30 s elapsed. The participants were instructed to find the best move for white as quickly as possible. Participants selected a move by dragging and dropping pieces in the diagram using the computer mouse. After participants had selected a move, feedback was presented for 2 s with respect to response latency (right-hand corner) and the move registered by the computer (left-hand corner). The next diagram was presented immediately after the feedback associated with the previous



Move	Elo	Response Time	*N*
a2-a6	2,116	16.2	35
e4-e8	1,791	12.1	132
d2-f4	1,883	20.7	16
d2-b4	1,770	24.7	2
f5-f6	1,666	23.2	2
e4-e7	1,741	22.2	2
e4-e5	1,344	27.3	1
b4-d2	1,882	11.3	1
d4-c2	2,317	3.3	1
e4-x0	1,448	6.3	1
None	1,959	30.0	22

Figure 2. Item 18 of the choose-a-move test A. Move a2–a6 is the correct answer, move e4–x0 means that the queen is accidentally dropped outside the board, and moves b4–d2 and d4–c2 are illegal

diagram. During the test, participants did not receive any information about the correctness of their moves. Responses were scored correct (best move) or incorrect (all other moves) and summed to construct test scores and subscale scores.

Motivation questionnaire

The motivation questionnaire was constructed based on the performance motivation test (Hermans, 1976). The questionnaire consisted of three sets of 10 items. Each set of items measured a separate motivation trait: positive fear of failure, negative fear of failure, and desire to win. Items were both positively and negatively worded and included statements such as “Overpowering my opponent makes me feel good.” Participants were asked to indicate, within 10 s, their level of agreement on a 5-point scale using the computer mouse. Test and subscale scores were calculated by summing the item scores (contraindicative items were rescored).

Predict-a-move test

This test is similar to the choose-a-move test. As in the choose-a-move test, each item visually represented a chess position, and participants were asked to indicate the move they believed to be the best within 30 s. In the choose-a-move test the items were completely unrelated, and the items were constructed in such a way that one move (i.e., the solution) was vastly superior to all other legal moves. In the predict-a-move test, however, the items were the subsequent moves for white from the grand master game Liss–Hector, Copenhagen 1996. Therefore, subsequent items were related, and for the majority of items no unique solution existed. For each item, every legal move for white was assigned points of merit on a scale from 0 to 5. We attempted to objectify the assignment process as much as possible by consulting chess experts, performing computer-aided analysis, and using a checklist of principles of good play, based mainly on Karpov and Mazukewitch (1987). We decided to give points not only for Karpov's seven rules of good play but also for the difficulty of the move (a forced exchange is easier to find than a three-move combination winning a pawn). To avoid discussion with opening theorists we decided to give 1 point for all moves known as playable for the entire opening stage.

In the predict-a-move test participants had to indicate, for each item, what they believed to be the best move for white. After participants had made their choice, the white move actually played by Liss and the black response move played by Hector were presented to the participant in brief succession. This procedure resulted in a new diagram that constituted the next item for the participant. After completing the final item, participants were asked to indicate whether they had encountered the Liss–Hector game before participating in the ACT.

Verbal knowledge questionnaire

The verbal knowledge questionnaire consisted of 15 four-alternative multiple-choice questions (adapted in part from Pfau & Murphy, 1988) that varied in difficulty. Participants were asked to select a response alternative with the computer mouse within 15 s. Four questions referred to opening knowledge, four questions referred to positional knowledge, and four questions referred to endgame knowledge. The remaining three questions referred to mental imagery, or the capacity to visualize chess moves. An example item for mental imagery is "What is the minimum number of moves required to transfer a knight from square f6 to square a1?"

Recall test

The recall test consisted of 18 items (for a comparable test, see Ericsson & Oliver, 1984). Each item involved the presentation of a chess diagram for 10 s, followed by a blank screen for 2 s. Next, an empty chess diagram (i.e., a diagram without any chess pieces) appeared, with one square marked by a black circle. The participant's task was to recall which chess piece had just occupied the marked square. The participant could select a chess piece from a list of chess pieces (including a symbol for an empty square) displayed to the right of the chess diagram. The participant was instructed to select a chess piece within 10 s using the computer mouse.

In the recall test, three factors were manipulated. The first factor was typicality of the chess position. As mentioned earlier, some controversy exists with respect to Simon and Chase's (1973) claim that the recall advantage of expert players is eliminated when chess positions are very atypical or random. We tested this claim by using chess positions in the recall test that were judged by chess experts to be very typical, not very typical, or random (i.e., random locations). The second and third factors of the recall test were exploratory and concerned the number of pieces (i.e., memory load, varied on three levels: 0–16 pieces, 17–24 pieces, or 25–32 pieces) and the position of the marked square (i.e., either a central square or a peripheral square).

EXPERIMENT

METHOD

Participants

Figure 1 shows the distribution of the Elo rating for the 259 participants. Participants were tested during the 1998 open Dutch championship in Dieren, the Netherlands. Of the 259 participants, 234 completed the entire ACT (i.e., all subtests); 215 of these 234 participants also had an official Elo rating, and a group of 159 participants completed the entire ACT, had an official Elo rating, and participated in the Dieren tournament. The latter group is important because participation in the tournament provided a second external criterion for the ACT (Elo rating being the first external criterion) because performance in chess tournaments is quantified by tournament performance rating (TPR). TPR is an Elo-based indication of chess-playing ability that is determined by their performance in a specific chess tournament.

Table 1 provides a list of characteristics of 215 participants who completed all tests and had an Elo rating. Almost all analyses are based on these data. Among the participants were 14 women (mean Elo = 1,648). We decided to include their results in our analyses because the Elo ratings for almost all of these women are based on club and tournament games against men and therefore probably are unbiased. Fifty participants also had a Fédération Internationale des Échecs (FIDE) rating. FIDE ratings correlate well with Dutch ratings (>.93 in our sample) but are on average about 50 points higher. Therefore, we decided to use the Dutch rating only.

Procedure

The 1998 Dieren tournament lasted 10 days. During the tournament, six computers were available for testing in a quiet corner of the tournament hall. The ACT was available in both Dutch and English. Completion of the ACT took about 1 hr. Participants were not rewarded financially but were given the opportunity to compare their scores with the distribution of scores from other participants. The subtests were administered in the following order: (1) the choose-a-move test part A, (2) the motivation questionnaire, (3) the predict-a-move test, (4) the verbal

Table 1. Participant characteristics

Group	Elo	TPR	Age	Women	Comp.	Hours	Games	*N*(tour)	*N*
Main	2,151 (157)	2,203 (182)	24 (8.4)	.02	.84	1.21 (0.90)	47 (30)	97	44
A	1,848 (96)	1,867 (158)	27 (13.8)	.00	.73	1.07 (1.06)	36 (24)	54	26
B	1,660 (102)	1,664 (152)	33 (17.3)	.03	.63	.96 (0.63)	29 (17)	75	30
C	1,474 (159)	1,518 (251)	36 (16.9)	.11	.71	1.01 (0.96)	25 (24)	58	18
6.1	1,741 (185)	1,722 (194)	31 (12.8)	.16	.77	.71 (0.58)	34 (19)	60	19
6.2	1,610 (81)	1,639 (201)	41 (18.4)	.00	.50	1.21 (0.99)	18 (16)	34	7
Rapid	1,748 (388)	1,808 (294)	24 (9.9)	.22	.88	1.44 (0.88)	37 (44)	117	9
G&F	2,027 (168)	— —	17 (1.4)	.17	.33	.75 (0.42)	49 (14)	31	6
Others	1,959 (273)	— —	36 (16.1)	.07	.68	1.07 (0.92)	47 (30)	—	56
Mean or total	1,865 (284)	1,851 (309)	31 (15)	.07	.72	1.06 (0.87)	35 (26)	526	215

Note. Standard deviations are given in parentheses. The tournament consisted of different groups. Groups “main,” A, B, and C have decreasing requirements with respect to skill level; groups 6.1 and 6.2 are amateurs playing a shorter tournament; “rapid” refers to a group of players participating in a subtournament with fast time controls; G&F (Glorney & Faber) refers to European youth team tournaments. Participants from the “rapid” group had very little opportunity to complete the ACT. Comp. = proportion of participants with computer experience; Hours = number of hours of daily chess practice; *N* = number of participants in each tournament group who completed the ACT; *N*(tour) = number of participants in the tournament groups; TPR = tournament performance rating; Women = proportion of women in the groups.

knowledge questionnaire, (5) the recall test, and (6) the choose-a-move test part B. This choice of order was motivated by the desire to present an optimal mixture of test formats. Pilot testing showed that the mental load of choose-a-move tests and predict-a-move tests was severe.

Before starting the ACT, participants were asked some general questions (i.e., name, sex, age, rating, nationality, hours of practice, number of games played last year, and computer experience). Participants' names were recorded to verify their Elo ratings and to obtain their TPRs. The names of the participants were later removed from the dataset. Participants were informed that their performance on the ACT would be anonymous. A general instruction was given that explained the purpose of the ACT. A short instruction and some training items preceded each subtest. Participants were told to respond accurately and quickly.

RESULTS

We first present the results pertaining to the characteristics of the tests such as reliability, scalability, relationship between accuracy and response time (RT), and factor structure. In the second part of the results section we investigate the validity of the tests with regression analyses on Elo rating and TPR. Finally, we test the effect of typicality on recall of chess positions. The computerized administration of the ACT did not lead to any serious complaints. Several participants did complain about the limited time available for completing the motivation questionnaire. Twenty-five participants did not complete the ACT because of obligations elsewhere because the total ACT takes an hour to complete. The incomplete data were discarded. We have no indication that incompleteness of data was related to any of the background variables. The ACT yielded a large amount of data (cf. Figure 2). The background variables sex ($r = .21$), age ($r = -.20$, ranging from 11 to 78 years), hours of practice per day ($r = .14$), and number of games played last year ($r = .29$) showed weak but significant correlations with Elo rating (see Ericsson & Lehmann, 1996, for a discussion of the last two background variables).

Scale Analysis

Reliability. Table 2 summarizes the main results of the scale analyses. The choose-a-move test A, with mean sum score of 19.2, $SD = 6.2$, was highly reliable, Cronbach's $\alpha = .87$. The three subscales of the choose-a-move test A (i.e., tactical, positional, and endgame items) were also sufficiently reliable in isolation, $.6 < \alpha < .8$. The sum score of the motivation questionnaire, $M = 98.3$, $SD = 11.9$, demonstrated acceptable reliability, $\alpha = .74$, but out of the three subscales, the third subscale (i.e., desire to win) was unreliable, $\alpha = .31$. The predict-a-move test, $M = 49.1$, $SD = 9.1$, and the verbal knowledge questionnaire, $M = 10.2$, $SD = 2.8$, were also fairly reliable, although the reliabilities of the subscales of the verbal knowledge questionnaire were low, $\alpha < .5$. The latter was also true for the recall test,

Table 2. Reliability and validity

		Accuracy			CISRT		
Test	*N*	α	i.i.r.	*r*(Elo)	α	i.i.r.	*r*(Elo)
Choose-a-move A	40	.87	.14	.77	.91	.19	.78
Tactical	20	.80	.15	.68	.85	.20	.71
Positional	10	.68	.17	.65	.73	.23	.69
Endgame	10	.60	.13	.66	.71	.18	.71
Motivation	30	.74	.09	.22	—	—	—
Positive fear	10	.63	.15	.19	—	—	—
Negative fear	10	.62	.14	.08	—	—	—
Desire to win	10	.31	.04	.19	—	—	—
Predict-a-move	42	.71	.09	.63	.88	.16	.53
Opening	12	.56	.13	.35	.79	.25	.33
Middle game	17	.48	.07	.52	.69	.11	.50
Endgame	13	.57	.12	.47	.76	.22	.49
Verbal knowledge	18	.67	.10	.55	.8	.17	.54
Opening	5	.41	.11	.4	.52	.18	.43
Middle game	5	.41	.14	.31	.64	.24	.37
Endgame	5	.33	.09	.47	.48	.15	.46
Imagery	3	.11	.04	.30	.24	.10	.45
Recall	18	.66	.10	.51	.74	.13	.50
Random	6	.26	.06	.25	.35	.18	.26
Low frequency	6	.53	.16	.43	.57	.18	.43
High frequency	6	.43	.11	.46	.50	.14	.48
Choose-a-move B	40	.90	.17	.81	.93	.23	.81
Tactical	20	.86	.23	.76	.90	.29	.77
Positional	10	.55	.11	.67	.65	.16	.67
Endgame	10	.67	.16	.66	.74	.21	.71

Note. Cronbach's α, mean interitem correlation (i.i.r.), and correlation with Elo rating are computed for two measures, sum score (accuracy) and sum score weighted with response time (correct item summed residual time, CISRT) for each test and each subtest. Because the motivation tests are not ability tests, CISRT is not computable.

$M = 9.9$, $SD = 3.18$. Finally, the choose-a-move test B, $M = 21.9$, $SD = 6.9$, was very reliable, $\alpha = .9$.

Alternative scorings of the predict-a-move test did not increase reliability or validity. For example, scoring 1 for good or playable moves and 0 for bad moves decreased the correlation with Elo to .60, but this decrease was not significant, $Z = 1.32$, $p > .05$. Therefore, the choice of scoring of response alternatives in the predict-a-move test probably is not particularly important. Only two participants claimed to know the game used in predict-a-move test, but their scores were well below average.

Two item response models for the ACT. In addition to the standard scale analyses we fitted a one-parameter and a two-parameter item response model (Lord, 1980) to the data of the choose-a-move tests A and B. In item response theory, items and people are placed on the same scale. The item response function determines the probability of a correct answer given the difference between difficulty of the item and ability of the person.[2] The one-parameter model (i.e., the Rasch model), allows items to differ in difficulty but not in discriminatory power. That is, all item functions are equal in form (i.e., discriminating power). The two-parameter model (i.e., the Birnbaum model) allows items to differ both in difficulty and in discriminatory power. We used the BILOG v1.1 program (Bock & Aitkin, 1981), which applies marginal maximum likelihood estimation. For both choose-a-move tests A and B, the one-parameter model did not fit the data according to the chi-square test, $\chi^2(155) = 272.2$, $p < .001$, and $\chi^2(148) = 281.1$, $p < .001$, respectively. The two-parameter model did fit the data of tests A and B, $\chi^2(139) = 109$, $p = .96$, and $\chi^2(141) = 154.2$, $p < .21$, respectively. We investigated whether the ability estimates derived from these item response models correlate more with Elo rating than does the simple test sum score measure. Contrary to expectation, the correlation with Elo rating decreased slightly when the ability estimates of the item response models were used instead of the accuracy sum score: For choose-a-move A, *r*(Elo, Rasch scores) = .75, *r*(Elo, Birnbaum scores) = .76; for choose-a-move B, *r*(Elo, Rasch scores) = .79, *r*(Elo, Birnbaum scores) = .80. Elimination of a few less reliable items did not improve the results. Therefore, the use of item response models instead of the accuracy sum score did not lead to an increase in external validity. Of course, such an increase in external validity can be expected in the case of adaptive testing (Wainer, 1990). Adaptive testing, in which the selection of test items is guided by the success rate on previous items, may reduce the number of items needed for a reliable test score.[3]

Combining response accuracy and RT. An important advantage of computerized testing is the automatic collection of RTs. In many tasks, response accuracy is determined to a large extent by the amount of time the participant invests in the problem at hand. The empirical phenomenon that participants of equal ability may show completely different behavior depending on whether they value speedy performance or accurate performance is called the speed–accuracy trade-off (Wickelgren, 1977).

This speed–accuracy trade-off also depends on age. In a regression analysis on RTs, with accuracy sum score and Elo rating as covariates, age is an important predictor, $B = .43$, $N = 215$, $t = 6.6$, $p < .001$. Older participants are significantly slower on choose-a-move tests. A complete evaluation of a participant's behavior on a test therefore entails a combination of information from response accuracy and response latency. Adequate use of RTs in computing test scores may increase the reliability and validity of a psychometric test (Dennis & Evans, 1996). We expected the benefits of incorporating RT to be particularly pronounced when response accuracy is uninformative (i.e., for easy items that almost every participant will solve correctly).

The correct and incorrect RTs are nonnormally distributed (i.e., skewed to the right), and the distribution of error RTs shows a peak just before the end of the time limit. The mean RT for incorrect responses did not correlate significantly with Elo rating, $r = -.02$, $N = 215$, $p > .05$ for choose-a-move A and $r = -.11$, $N = 215$, $p > .05$ for choose-a-move B, whereas mean RT on correct responses correlated significantly with Elo rating, $r = -.30$, $N = 215$, $p < .001$ and $r = -.26$, $N = 215$, $p < .001$, respectively. These significant negative correlations reinforce the idea that prediction of Elo rating could be improved when both response accuracy and response latency are incorporated in the scoring of the ACT.

In the following, we will evaluate three different measures of test performance that address the speed–accuracy trade-off. Dennis and Evans (1996) studied two measures that combine information from response latency and response accuracy for multiple-choice items such as those used in the ACT. The first measure is the ratio index (RI), defined as RI = $p/\bar{t}$, where p stands for average probability of correct responding and $\bar{t}$ denotes average RT. The second measure discussed by Dennis and Evans is the log A index (LAI), defined as LAI = $-1/(\bar{t} - t_{\mathrm{min}})\ln([A - \mathrm{logit}(p)]/A)$, where t_{min} denotes the minimal RT to respond above chance and A denotes asymptotic accuracy.

A third measure of speed–accuracy test performance, introduced here, is called correct item summed residual time (CISRT). All ACT items have a maximum completion time MT (e.g., $MT = 30$ s for the choose-a-move items); the CISRT score depends on how much time is left of the maximum time allowed if and only if the response is correct. If the response is incorrect, the CISRT score is zero. Thus, CISRT = $\Sigma_i \mathrm{Acc}_i(MT - t_i)$, where i is the item index, Acc denotes response accuracy (0 for incorrect and 1 for correct), and t denotes response latency. In other words, the CISRT is obtained by summing the residual time for each item that was completed correctly. An advantage of CISRT is that it can be computed separately for each item.

Table 3 shows that the correlation between test performance quantified by CISRT and Elo rating was somewhat higher than the correlation between response accuracy and Elo rating in the upper half of the Elo range. This difference in correlation is significant only for choose-a-move test A, $Z = 1.69$, $p < .05$, according to a test proposed by Meng, Rosenthal, and Rubin (1992). In contrast, the correlations between the two other speed–accuracy measures of test performance (i.e., the RI and LAI) and Elo rating were lower than when accuracy only was used to quantify test performance.

Table 3. Correlation between speed–accuracy measures and Elo rating for the choose-a-move tests

	Choose-a-move A			Choose-a-move B		
		Elo	Elo		Elo	Elo
	Elo	< 1,852	≥ 1,852	Elo	< 1,852	≥ 1,852
Accuracy	.78*	.44**	.57**	.81**	.54**	.66**
RT	–.39**	.06	–.46**	–.32**	.54**	–.35**
CISRT	.79**	.38**	.65**	.81**	.50**	.70**
RI	.69**	.30**	.57**	.69**	.38**	.59**
LAI	.50**	.38**	.44**	.52**	.43**	.50**

Note. CISRT = correct item summed residual time; LAI = log A index (Dennis & Evans, 1996); RI = ratio index; RT = response time. CISRT predicts Elo rating better in high than in low Elo rating groups (median split), possibly because CISRT performs better on easy items (see Figure 3).
*$p < .01$. **$p < .001$.

The advantage of using the CISRT is illustrated more convincingly on the item level. The difference between simple response accuracy (correct/incorrect) and CISRT in correlation with Elo over the 80 items of the choose-a-move tests is significant, mean difference in correlation = .11, $t = 2.18$, $df = 157.9$, $p < .05$. This difference strongly correlates with the difficulty (i.e., number of correct answers) of the items, $r = .78$, $N = 80$, $p < .001$, indicating that the advantage of using the speed–accuracy CISRT measure was more pronounced for the easy items (Figure 3).

Misleading items. We discuss one other use of RT, regarding misleading items. Items can be difficult for a number of reasons (e.g., interference, depth of combination), and some items are difficult because there is an attractive but wrong move. RTs can be helpful in objectively assessing characteristics such as misleadingness. To illustrate, we quantify misleadingness in two ways.

First, we compute how often the most favorite wrong answer is chosen (actually, we compute the quotient of this number and the total number of errors on this item). So, for item 18 (Figure 2) this quotient is .73, that is, 132 choices of e4–e8 divided by the total number of errors, which is 215 – 35 = 180. In comparison to other items, this quotient is high, suggesting that this wrong move was particularly attractive.

Second, we compute the difference between the mean RT of the wrong responses with the mean RT of the correct responses. We expect that at misleading items the wrong response is selected quickly, whereas the selection of the correct response takes more time (because the incorrect but attractive alternative is evaluated and rejected first). In the case of item 18 this difference was close to zero, .80 s, whereas on most items correct

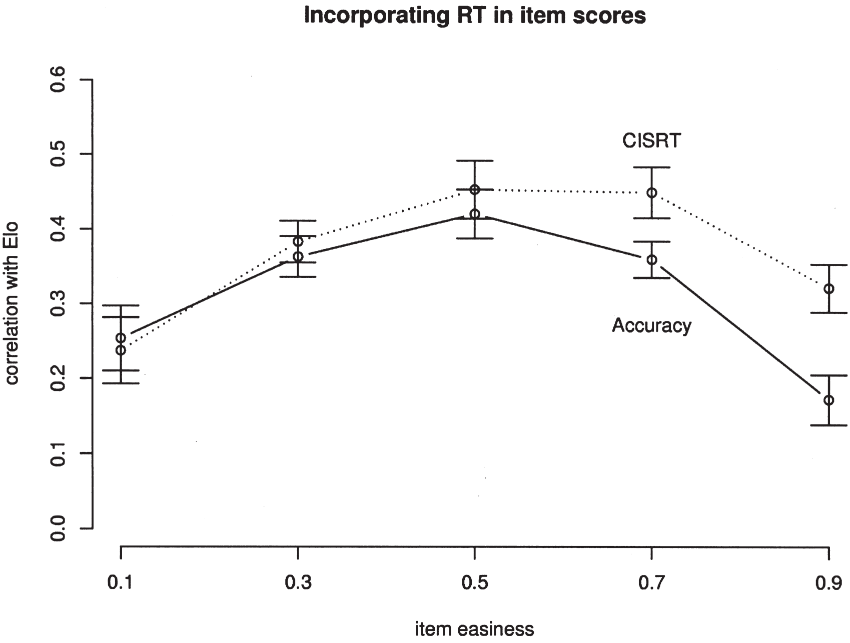


Figure 3. Correct item summed residual time (CISRT) score compared with the accuracy-only score for all items of choose-a-move tests A and B

responses are much faster than incorrect responses, mean difference of –4.7 s, $t(39) = -6.9$, $p < .001$.

The item estimates of misleadingness of these two methods had a correlation of .53, $p < .001$, for the choose-a-move A test. This significant correlation between two completely different measures suggests that some items are indeed misleading, with item 18 of choose-a-move test A as a good example. A similar effect could be demonstrated for choose-a-move test B.

Factor structure. To further examine the test characteristics of the ACT, we investigated the relationships between the subtests with factor analytic methods. An exploratory factor analysis using the multivariate analysis (MVA) package of the *R* program (Ihaka & Gentleman, 1996) for the six tests of the ACT resulted in a one-factor model that fitted the data, $\chi^2(9) = 14.2$, $p = .12$. Subsequently, to further gain insight in the structure of the ACT, the scores on all 19 subscales were analyzed with exploratory maximum likelihood factor analysis. Four factors were sufficient to account for the correlation matrix, $\chi^2(101) = 103.02$, $p = .42$. Table 4 shows the Promax (Hendrickson & White, 1964) rotated factor structure of the ACT subscales. As expected from the high correlations between the ACT scales, this Promax rotated solution was much more interpretable than the orthogonal solution. The first factor explained 21% of the variance and could be interpreted as the main chess factor on which the choose-a-move tests A and B both load high. The second factor explained 8% of the variance and is related primarily to memory. The first two subscales of the predict-a-move test, opening and middle game, load on this factor. Memory might influence performance in the predict-a-move test because the items are presented in chronological order, and recall of the specific opening line (Ruy Lopez) might play a role. Memory is the only factor that correlates significantly with age, $r = -.36$, $N = 215$, $p < .001$. The third factor, explaining 7% of the variance, was associated primarily with motivation, whereas the fourth factor, explaining 7% of the variance, related mainly to verbal knowledge, particularly to opening knowledge.

Table 4. Promax rotated factor solution of the subscales of the Amsterdam Chess Test

	Factor 1	Factor 2	Factor 3	Factor 4
Choose-a-move A				
Tactical	0.74			
Positional	0.69			
Endgame	0.84		–0.19	
Motivation				
Positive fear	0.28		0.48	
Negative fear			0.73	
Desire to win			0.72	
Predict-a-move				
Opening		0.44		
Middle game	0.21	0.36		
Endgame	0.49			
Verbal knowledge				
Opening				1.06
Middle game	0.20			0.33
Endgame	0.50			
Imagery	0.22			
Recall				
Random		0.61		
Low frequency		0.66		
High frequency	0.16	0.36		
Choose-a-move B				
Tactical	0.66	0.25		
Positional	0.59	0.17		
Endgame	0.81			

Note. Factor loadings with absolute values smaller than 0.15 are not listed. In Promax rotated solutions, loadings are interpreted as regression coefficients, and loadings higher than 1 are possible.

Confirmatory and exploratory factor analyses were used to study the factor structure of the subtests of the choose-a-move tests A and B. A confirmatory three-factor model (using structural equation LISREL modeling, Jöreskog & Sorbom, 1996) with separate factors for tactical, positional, and endgame ability fitted the data well, $\chi^2(6) = 6.55$, $p = .37$. However, because the factors of the three-factor model were highly correlated, a simpler one-factor model with just one correlated residual (between the tactical subtests) also fitted the data well, $\chi^2(8) = 8.47$, $p = .39$. These high factor correlations complicate any attempt to test the hypothesis that differences between players in the higher Elo rating ranges are related more to positional and endgame knowledge than to tactical ability. Moreover, correlations with Elo were very sensitive to how the Elo ranges were chosen. To conclude, we were not able to find convincing evidence for this hypothesis.

Validity. The present study yielded two related ($r = .88$) measures, Elo rating and TPR, that can be used to assess the criterion validity of the ACT. Elo ratings are established on the basis of many more games than were used for the TPR and therefore probably are more reliable. On the other hand, Elo ratings are considered to be conservative and stagnant (Sonas, 2002). Sonas showed that a more dynamic rating formula improves the prediction of chess results.

We performed two sets of regression analyses to investigate the contribution of the different tests to the prediction of Elo rating and TPR. First, the chess tests and several background variables were used to explain the variance in Elo rating. Second, we tested whether the chess tests could still add to the explanation of TPR after including Elo rating in the set of predictors.

Predicting Elo. The correlations with Elo are highest for the sum scores of the choose-a-move tests (.77 and .81; see Table 2) and compare well with the correlations reported by Pfau and Murphy (1988). The regression equation predicting Elo from choose-a-move A is 1,189 (40.3) + 35.15 (2.0) × number correct ($SE = 181.5$). The regression equation predicting Elo from choose-a-move B is 1,169 (36.5) + 33.10 (1.6) × number correct ($SE = 167.2$).

Table 2 shows the correlation with Elo rating for each subtest of the ACT. The top panel of Figure 4 shows that criterion validity was still high even within specific ranges of the Elo rating scale. Thus, the ACT is also valid to measure expertise differences within the group of chess experts.

Next we performed stepwise regression analyses using the six test scores of the ACT as predictors of Elo. The parameter estimates are shown in the upper left half of Table 5. Only the choose-a-move and the predict-a-move tests were entered in the regression explaining 70% of the variance in Elo. However, the exclusion of some of the ACT subtests in the regression

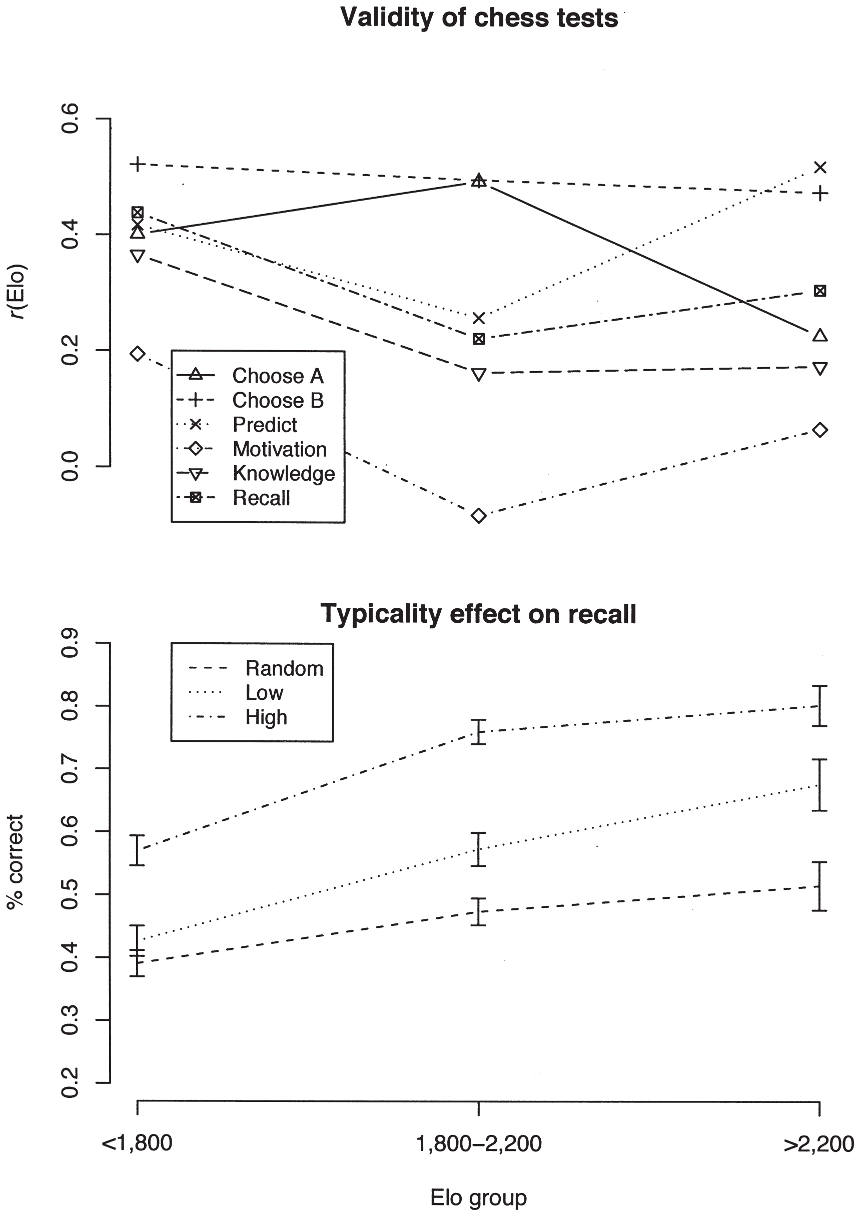


Figure 4. Validity of chess tests and typicality effect on recall within 3 Elo ranges (<1,800, *N* = 93; 1,800–2,200, *N* = 98; >2,200, *N* = 33)

Table 5. Parameter estimates of regression models of Elo rating and TPR

	Dependent: Elo				Dependent: TPR			
	Estimate	*SE*	*B*	*t* value	Estimate	*SE*	*B*	*t* value
Intercept	958.8	36.40		16.25***	−120.00	113.6		−1.06
Elo					0.78	.06	.71	13.50***
Choose A	10.5	3.42	.23	3.06**	9.16	2.67	.18	3.44**
Motivation					3.78	.96	.14	3.94***
Predict	5.5	1.53	.18	3.59***				
Verbal								
Recall								
Choose B	20.7	3.03	.51	6.83***				
Intercept	1,810.9	70.2		25.79***	501.2	166.3		3.01**
Elo					0.8	0.1	0.70	9.48***
Factor 1	265.1	12.5	1.20	21.22***	69.9	25.7	0.29	2.72**
Factor 2	206.3	14.9	0.69	13.86***	55.7	23.9	0.17	2.33*
Factor 3	84.5	13.1	0.26	6.44***	57.9	16.4	0.16	3.53***
Factor 4	142.7	11.9	0.59	12.04***	33.8	17.3	0.13	1.95
Comp.	23.0	24.7	0.04	0.93	−27.1	27.7	−0.04	−0.98
Sex	−110.8	51.0	−0.08	−2.17*	59.5	64.2	0.04	0.93
Age	3.9	0.9	0.19	4.32***	−1.7	1.1	−0.07	−1.52
Hours	−6.8	13.5	−0.02	−0.50	15.3	15.1	0.04	1.01
Games	1.0	0.4	0.09	2.35*	−0.8	0.5	−0.07	−1.73

Note. The upper part of the table displays the results of stepwise regression analyses on Elo and tournament performance rating (TPR). The lower part shows the simultaneous entry regression analyses with factor scores instead of test scores. In these latter analyses the background variables were also included. Factors 1 to 4 were associated respectively with choose-a-move, recall, motivation, and verbal knowledge (see Table 4). Comp. = computer experience (1 = yes, 2 = no); games = number of official games played the year before the Dieren 1998 tournament; hours = estimated number of hours devoted to chess per week; sex (1 = male, 2 = female).

*$p < .05$. **$p < .01$. ***$p < .001$.

equation for Elo and TPR should be interpreted with great care because the high correlations between the tests lead to collinearity.

According to the criteria of Belsley, Kuh, and Welsch (1980), problems with collinearity occurred for three of the six tests (e.g., conditioning index larger than 15). We decided to address the problem of collinearity by using the factor scores based on the Promax rotated factor solution instead of the raw test scores (cf. Table 4). The correlations between the factor scores are much lower, and the maximal conditioning index for these factors is 2.5.

In the lower left part of Table 5 we present a regression model that used the factor scores instead of the test scores. In these analyses we also included the predictors age, sex, computer experience, number of games played in the year before the Dieren 1998 tournament (July 1997–June 1998), and estimated practice hours per week. This regression model explained 75% of the variance in Elo rating. All four Promax factors contributed significantly to the prediction of Elo rating. The estimates for age, sex, and number of games also deviated significantly from zero. Both young participants and female participants, $N = 14$, had lower Elo ratings than would be expected on basis of their test scores. The standard error of estimate in this analysis was 147.4, 95% CI ≈ ±62, 95% PI ≈ ±292. The mean standard error of the predicted values (an estimate of the standard deviation of the average value of the dependent variable for cases that have the same values of the independent variables) was 31.7, $SD = 9.06$, which compares well with the estimated reliability of Elo ratings (between 25 and 54).

Predicting TPR. In the stepwise regression analysis for TPR, the predictors Elo, the choose-a-move test A, and the motivation test were entered as variables explaining 82% of the variance in TPR (see Table 5, upper right panel).

The full model (Table 5, lower right panel) explained 83% of the variance in TPR. The background variables did not add to the prediction of TPR, but the four chess factors did.[4] Excluding the background variables led to an explained variance of 82% (compared with 77% explained variance for a model with only Elo as predictor). Because variables other than Elo are able to explain a significant part of the variance in TPR, the difference between Elo and TPR cannot be completely attributed to the unreliability of TPRs. This supports the view, shared by many chess players, that Elo ratings lag behind real ability.

We can estimate this lag with the nonstandardized estimate for age in the model for Elo, which is 3.9 (Table 5). A regression analysis with TPR as the dependent variable and Elo and age as independent variables gives an estimate of −3.63, $p < .001$, for age. These estimates suggest that a player's Elo is about 3.5–4.0 points per year too low compared with older players. This effect might also be important for the discussion of age effects on chess skill (Charness, 1991).

Predictive validity. We were able to collect the Elo ratings of most participants before and after ACT administration. Figure 5 shows the correlation of the ACT tasks with Elo rating over the past 10 years (from 1993 to 2002, average $N = 175$). The vertical line gives the time of administration of the ACT in 1998. Of course, the Elo rating autocorrelation decreases as lag increases. However, the correlation between Elo rating and several subtests of the ACT increased for some time and then became stable. This increase in correlation was statistically significant for choose-a-move test

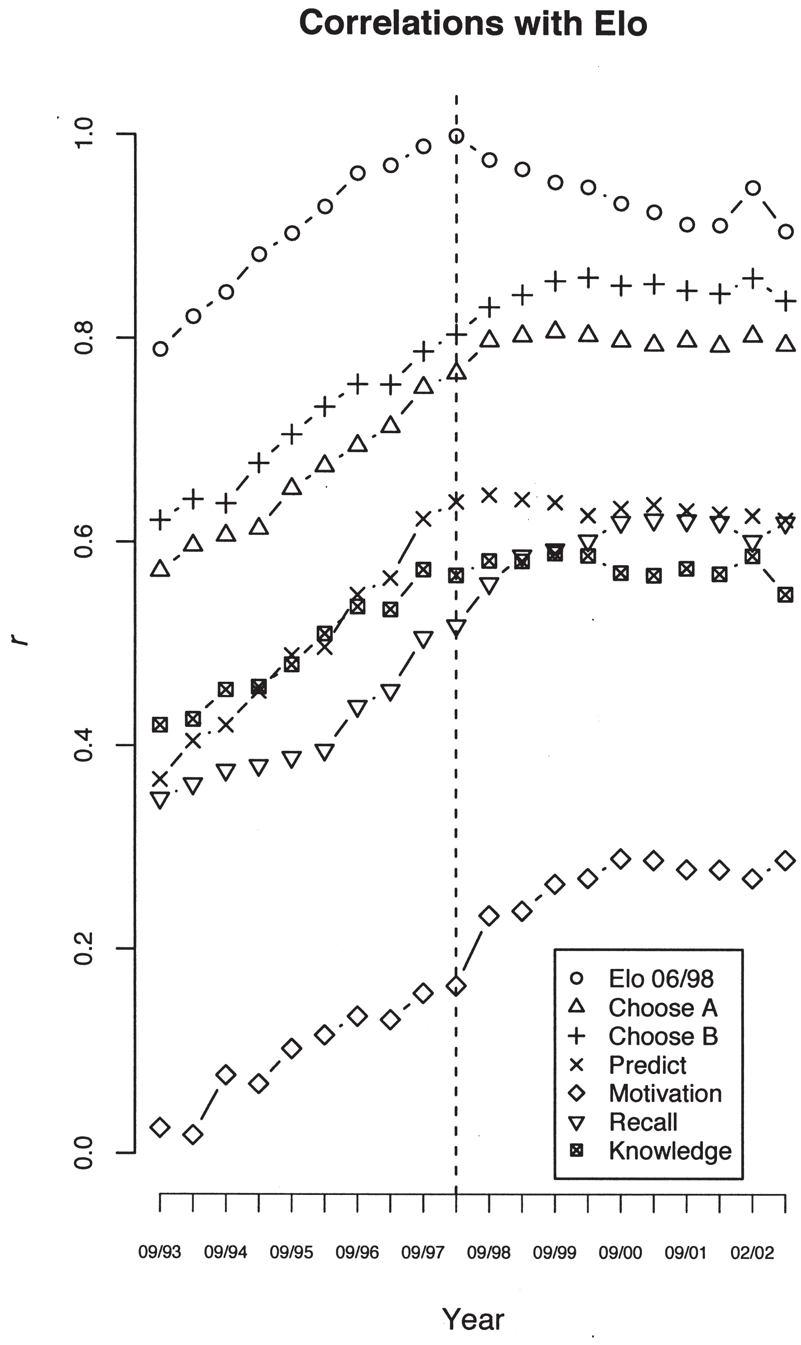


Figure 5. Increase in predictive validity in the years after the administration of the ACT in June 1998

A and the motivation test in the first half year and significant for choose-a-move test B and the recall test the first year and a half (according to the test for correlated correlations proposed by Meng et al., 1992). The correlations of Elo rating with choose-a-move test B, the recall test, and the motivation test were still significantly larger in September 2002 than in May 1998, the time of the administration of these tests (i.e., Z values of 4.4, 1.8, and 3.4, respectively). This pattern in predictive validity is noteworthy and highlights the validity and usefulness of the ACT.

Typicality effect on recall

The recall test consisted of 18 items that varied on three levels of typicality (high, low, and random), three levels of number of pieces (0–16, 17–24, and 25–32), and two levels of position of marked square (central and peripheral). The sum score of this test correlated .51 with Elo rating, which is significantly lower than the correlation between choose-a-move test A and Elo, $Z = 6.0$, $p < .001$. We focus here on tests of the hypothesis that the effect of ability on recall performance decreases as typicality decreases.[5]

An ANOVA with number correct as a dependent measure and typicality (three levels) as a within-factor measure and Elo rating as a covariate yielded a significant main effect of typicality, $F(2, 426) = 3.16$, $p < .05$; an effect of the covariate, $F(1, 213) = 73.5$, $p < .001$; and a significant interaction between typicality and the Elo covariate, $F(2, 426) = 7.47$, $p < .01$. The lower panel of Figure 4 illustrates the effects. It is clearly not true that all chess players are equally poor in recall of random positions. The higher Elo rating groups performed much better on the random items than did the lower Elo rating groups, $F(3, 6) = 4.41$, $p < .01$. This finding is inconsistent with that of Simon and Chase (1973) but is consistent with later work (Gobet & Simon, 1996).

Yet there is also a significant interaction of typicality and Elo. The correlation between Elo rating and recall is lower for random positions, $r = .25$, $t(213) = 3.84$, $p < .001$, than it is for low-frequency positions, $r = .43$, or high-frequency positions, $r = .46$, $z = -2.48$ and $z = -2.98$, respectively.

DISCUSSION

Analyses of the ACT data have answered most questions posed at the beginning of this article. First of all, it can be safely concluded that it is possible to obtain a valid and reliable measure of chess expertise in a limited amount of time. This can be accomplished using an IQ-like task composed of chess problems that can be found in chess books, chess magazines, and newspapers. The findings reported here show that choose-a-move tests can provide an accurate assessment of chess expertise in about 15 min. Because a reliable Elo rating requires at least 20 serious chess games, most of which take several hours, such a quick assessment is a convincing illustration of the value of psychological testing. The regression models showed that performance on the ACT explained 75% of the variance in Elo rating. Note that part of the 25% unexplained variance is due to the unreliability (lag) of our validity criterion Elo rating. It is remarkable that the ACT explained a significant additional amount of the variance (5%) in TPR when Elo rating was included in the set of predictors.[6] Furthermore, the predictive power of the ACT increased until a year and a half after the assessment. The results imply that chess trainers and tournament organizers can safely use this type of test to assess the effect of training, to find weak points of players, and to assign players to training and tournament groups.

A possible point of criticism concerns the short time limits used in the ACT. We expected that it was more efficient to use many items with short time limits than to use just a few items with longer time limits, but we did not test this expectation. On one hand, there is ample evidence that chess players use different mental capacities when thinking many minutes over one chess move than when they play a entire blitz game in just a few minutes (Chabris & Hearst, 2003). On the other hand, correlations between Elo ratings and blitz ratings usually are very high (Burns, 2004). Burns showed that up to 81% of the chess skill variance (measured by rating) was accounted for by how players perform with less than 5% of the normal time available. Therefore, we did not expect a substantial validity increase when longer time limits were used. Finally, results of Calderwood, Klein, and Crandall (1988) suggest that skill differences are amplified under time pressure. However, this result is based on data from only six participants.

Another issue concerns the use of computers. Computer experience was not a significant predictor of Elo rating in any of our analyses, but the measure of computer experience consisted of only two categories (yes or no). A more advanced measurement might change the picture, although we do not think that computer experience explains the independent contribution of age in the regression analyses (cf. Charness, 1981b). Age appeared to be an important variable in many of our analyses. First, older participants are significantly slower on the choose-a-move tests, also when corrected for Elo rating and test score. Second, age correlates strongly with the recall factor. Third, age is an important predictor in the regression analysis on Elo, much more important than the other background variables. Finally, in predicting TPR from Elo rating and age, age appeared to be an important predictor.

We investigated several test formats other than the choose-a-move test format. The predict-a-move test might provide a higher external validity, but this could not be demonstrated. This test was less reliable and less valid than the choose-a-move test. However, the predict-a-move test contributed independently to the prediction of Elo rating and the predict-a-move subscales loaded on the choose-a-move and recall factors.

The contribution of the verbal knowledge test is limited, although some evidence of an independent contribution to the prediction of both Elo rating and TPR was present in the regression analyses using the factor scores. The ACT verbal knowledge test was also rather short and is open to improvement. Pfau and Murphy (1988) used a much longer test and found a correlation of .69, which is higher than the .55 correlation of the ACT verbal knowledge test.

The results of the motivation test are mixed. The correlation with Elo rating is fairly low, but the regression models clearly show an independent contribution in predicting chess expertise. Surprisingly, the predictive validity of the motivation questionnaire increases markedly after testing (Figure 5).

The correlation between the recall test and Elo was much lower than between the choose-a-move tests and Elo. A stepwise regression procedure led to the exclusion of the recall test. In the factor analysis, however, a recall factor could be identified. This recall factor significantly contributed to the explanation of Elo and TPR in the simultaneous entry regression analyses. Nonetheless, in simply predicting Elo the choose-a-move format is clearly superior to the recall test.

However, many subtests of the ACT differ in reliability and correlate highly. Because of collinearity the results of the stepwise regression analyses must be interpreted with care. The analyses using Promax rotated factor scores are somewhat more complicated but give a better idea of the relationship between the ACT, Elo, and TPR.

A more detailed analysis of the subscales of the choose-a-move tests showed that it is very difficult to detect differences in abilities associated with tactical, positional, or endgame items. The factors associated with these abilities correlate very highly. The fact that top players seem to be better than amateurs in all respects makes a case for a general ability factor in chess, much like the "*g*" factor in intelligence.

In addition to different test formats, we investigated different psychometric methods. Using item response models instead of the simple accuracy sum score did not increase the correlation with Elo rating. Increasing validity may not be the first aim of these models. Adaptive testing of chess ability using item response models might be an attractive future extension of the ACT.

Test performance on the ACT was quantified in the first place by response accuracy. However, response speed also proved to be indicative of chess proficiency, especially for easy items (cf. Figure 3). In the future, it would be better to inform respondents of the CISRT formula so that they may take this scoring rule into account. An important advantage of such a procedure is that it might solve the speed–accuracy problem for this type of achievement task by letting respondents know how many points they get for speed relative to accuracy so they can choose an optimal trade-off (Wickelgren, 1977).

We also showed that RTs yield information about the extent to which items can be misleading. Similar analyses can be performed to investigate difference between items that depend on factual knowledge (i.e., the ending of king and two knights against king alone usually is drawn) and items that require the detailed computation of long lines.

A final example of the usefulness of assessing response latency concerns the estimation of the probability of a correct guess (the guess parameter in three-parameter item response models). Inspection of the RTs on the choose-a-move items shows that the frequency of responses increased in the last 5 s. Participants tended to guess just before their time was up. The probability of a correct guess perhaps can be estimated by the relative frequency of correct responses in the last 5 s (these estimates are .12 and .14 for items of choose-a-move A and B, respectively). These estimates might be very useful in the application of item response models, for instance in estimating the guess parameter of the three-parameter item response theory model.

The recall test was also used to test what is perhaps the most famous prediction in chess research. Simon and Chase (1973) argued that the superiority of chess masters on chess memory tasks disappears when random positions are used. Others (e.g., Charness, 1981a, 1981b; Gobet & Simon, 1996) questioned the results of Simon and Chase. Our results reinforce these doubts. Experts clearly are superior to novices even when random positions are used. On the other hand, the observed interaction between typicality and Elo rating demonstrates that the superiority of experts in the recall task is more pronounced when regular, chesslike positions are used. There is no standard in the literature for presentation time. Presentation times varying between 1 s and 1 min have been applied. Results of Gobet and Simon (2000) and McGregor and Howes (2002) suggest that recall effects are robust for changes in presentation.

We distinguish three areas for future research. First, the ACT can be improved. The time limits for the motivation test and the verbal knowledge may have been too short. Several items of the recall test should be revised. The choose-a-move tests should include more very easy and very difficult items. We would especially like to try adaptive testing with methods based on item response theory. The wide range of differences in chess expertise between chess players (from real beginners to top grand masters) and the availability of thousands of chess problems provide excellent conditions for successful adaptive testing.

Second, in line with the last point, various psychometric techniques can be validated through chess testing research. We demonstrated the possibilities of using RTs in assessing chess ability and in testing specific hypotheses about items. This is important because the psychometric tradition has generally failed to confront the implications of the speed–accuracy trade-off for the way items or tests are scored (Dennis & Evans, 1996). Because computerized testing automatically yields the RTs, different solutions to the speed–accuracy trade-off may be investigated (Verstralen, Verhelst, & Bechger, 2001). To give just one example of an important question, What happens with the item difficulties and item reliabilities when the time for responding is changed from 30 s to 15 s?

Finally, we expect that the choose-a-move test format will enable the investigation of theories of the thought processes involved in chess playing. Content-specific theories about these thought processes could be tested by specially constructed chess problems. For instance, an ACT-like test could be used to measure the relative difficulty of certain types of combinations and the importance of the depth of a combination (Saariluoma, 1995). The application of ACT and similar test formats to theory testing is an interesting challenge for future work.

Notes

The authors are indebted to the organizing committee of the 1998 open Dutch chess championship in Dieren and to the Royal Dutch Chess Federation and Patrick Razenberg in particular. We thank Sandra de Blécourt and Jaap Kamminga for their help in constructing tests and collecting the data, Harry Vorst for constructing the motivation questionnaire, Conor Dolan for assisting in the LISREL analysis, and Willemijn Roorda for assisting in the RT analysis. We also thank Karel van der Weide (international master) and Dimitri Reinderman (international grand master) for checking the test items and making valuable suggestions for improvement. We thank Neil Charness, Anders Ericsson, and Fernand Gobet for their valuable comments on a previous draft of this article. More detailed information about the ACT and the test results is available from http://users.fmg.uva.nl/hvandermaas/chesshtml/act.htm.

Correspondence about this article should be addressed to Han L. J. van der Maas, Department of Psychology, University of Amsterdam, Roetersstraat 15, 1018 WB Amsterdam, The Netherlands (e-mail: hvandermaas@fmg.uva.nl). Received for publication September 2, 2003; revision received April 9, 2004.

1. The Internet search engine PsychINFO lists 175 articles in the past decade that have *chess* in their abstracts.

2. In item response theory this scale, used for item difficulties and person abilities, has an arbitrary mean and variance. There are many ways to fix the scale. In the case of a chess test it is a good idea to use the Elo scale. Item difficulties can then be expressed in terms of Elo rating with a straightforward interpretation: It determines the probability of win (correct solution within 30 s) of a player with a

certain Elo rating. For instance, the item in Figure 2 has a rating of 2,134. In fact, the Elo system and Rasch model share many elements (de Blécourt, 1998).

3. Internet chess servers such as www.chessclub.com could be used to investigate this form of testing. Each day thousand of players log on to play games and watch tournaments. The www.chessclub.com chess server also hosts bots, such as the "trainingbot," whose interface is very suitable for a computerized adaptive chess test. Adaptive chess tests may make it possible to measure small differences in chess proficiency in the extreme tails of the Elo rating distribution.

4. Although there were some extreme differences between Elo and TPR, excluding these outliers did not change the significance of the predictors.

5. Analyses of the other effects led to results that we were unable to interpret. The two low-frequency items with 0–16 pieces were unexpectedly very difficult (cf., Gobet et al., 2004). The effect of location (central versus peripheral) was masked by a possibly more important effect of relevance of square.

6. Note that the correlation between Elo rating and TPR is .88. This high correlation also results from the fact that the tournament group to which chess players are assigned restricts TPRs. This assignment to groups takes place on the basis of Elo rating.

References

Amidzic, O., Riehle, H.-J., Fehr, T., Wienbruch, C., & Elbert, T. (2001). Pattern of focal gamma-bursts in chess players. *Nature, 412,* 603.

Atherton, M., Zhuang, J., Bart, W. M., Hu, X., & He, S. (2003). A functional MRI study of high-level cognition. I. The game of chess. *Cognitive Brain Research, 16,* 26–31.

Batchelder, W. H., & Bershad, N. J. (1979). The statistical analysis of a Thurstonian model for rating chess players. *Journal of Mathematical Psychology, 19,* 39–60.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics.* New York: Wiley.

Binet, A. (1966). Mnemonic virtuosity: A study of chess players. *Genetic Psychological Monographs, 74,* 127–164. (Original work published 1983)

Bloch, M. (1994). *The art of combination.* London: International Chess Enterprises.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of the EM algorithm. *Psychometrika, 46,* 443–459.

Burns, B. D. (2004). The effects of speed on skilled chess performance. *Psychological Science, 15,* 442–447.

Calderwood, B., Klein, G. A., & Crandall, B. W. (1988). Time pressure, skill, and move quality in chess. *American Journal of Psychology, 101,* 481–493.

Chabris, C. F, & Hearst, E. S. (2003). Visualization, pattern recognition, and forward search: Effects of playing speed and sight of the position on grandmaster chess errors. *Cognitive Science, 27,* 637–648.

Charness, N. (1981a). Aging and skilled problem solving. *Journal of Experimental Psychology: General, 110,* 21–38.

Charness, N. (1981b). Search in chess: Aging and skill differences. *Journal of Experimental Psychology: Human Perception and Performance, 2,* 467–476.

Charness, N. (1991). Knowledge and search in chess. In K. A. Ericsson & J. Smith (Eds.), *Towards a general theory of expertise* (pp. 39–63). Cambridge: Cambridge University Press.

Charness, N., Reingold, E. M., Pomplun, M., & Stampe, D. M. (2001). The perceptual aspect of skilled performance in chess: Evidence from eye movements. *Memory & Cognition, 29,* 1146–1152.

Chase, W. G., & Simon, H. A. (1973). The mind's eye in chess. In W. G. Chase (Ed.), *Visual information processing* (pp. 215–281). New York: Academic Press.

de Blécourt, S. (1998). *The legacy of Arpad Elo.* Unpublished manuscript available at http://www.schaakbond.nl/rating/

de Groot, A. D. (1978). *Thought and choice in chess.* The Hague: Mouton. (Original work published 1946)

de Groot, A. D. (1986). Intuition in chess. *International Computer Chess Association Journal, 9,* 67–75.

de Groot, A. D., & Gobet, F. (1996). *Perception and memory in chess.* Assen, The Netherlands: Van Gorcum.

Dennis, I., & Evans, J. (1996). The speed–error trade-off problem in psychometric testing. *British Journal of Psychology, 87,* 105–129.

Djakov, I. N., Petrovsky, N. B., & Rudik, P. A. (1926). *Psihologia shakhmatnoi igry* [Chess psychology]. Moscow: Avtorov.

Elo, A. (1978). *The rating of chess players, past and present.* London: Batsford.

Ericsson, K. A. (2003). Development of elite performance and deliberate practice: An update from the perspective of the expert performance approach. In J. L. Starkes & K. A. Ericsson (Eds.), *Expert performance in sports: Advances in research on sport expertise* (pp. 50–83). Champaign, IL: Human Kinetics.

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review, 102,* 211–245.

Ericsson, K. A., & Lehmann, A. C. (1996). Expert and exceptional performance: Evidence of maximal adaptation to task constraints. *Annual Review of Psychology, 47,* 273–305.

Ericsson, K. A., & Oliver, W. (1984). *Skilled memory in blindfolded chess.* Paper presented at the annual meeting of the Psychonomic Society, San Antonio, TX.

Ericsson, K. A., Patel, V., & Kintsch, W. (2000). How experts' adaptations to representative task demands account for the expertise effect in memory recall: Comment on Vicente and Wang (1998). *Psychological Review, 107,* 578–592.

Ericsson, K. A., & Smith, J. (1991). Prospects and limits of the empirical study of expertise: An introduction. In K. A. Ericsson & J. Smith (Eds.), *Towards a general theory of expertise* (pp. 39–63). Cambridge: Cambridge University Press.

Gobet, F. (1997). A pattern-recognition theory of search in expert problem solving. *Thinking and Reasoning, 3,* 291–313.

Gobet, F. (1998). Expert memory: Comparison of four theories. *Cognition, 66,* 115–152.

Gobet, F. (2000). Some shortcomings of long-term working memory. *British Journal of Psychology, 91*(4), 551–570.

Gobet, F., Campitelli, G., & Waters, A. J. (2002). Rise of human intelligence: Comments on Howard (1999). *Intelligence, 30,* 303–311.

Gobet, F., De Voogt, A. J., & Retschitzki, J. (2004). *Moves in mind: The psychology of board games.* Hove, UK: Psychology Press.

Gobet, F., & Simon, H. A. (1996). Recall of rapidly presented random chess positions is a function of skill. *Psychonomic Bulletin & Review, 3,* 159–163.

Gobet, F., & Simon, H. A. (2000). Five seconds or sixty? Presentation time in expert memory. *Cognitive Science, 24,* 651–682.

Gobet, F., & Waters, A. J. (2003). The role of constraints in expert memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 1082–1094.

Hendrickson, A. E., & White, P. O. (1964). Promax: A quick method for rotation to orthogonal oblique structure. *British Journal of Statistical Psychology, 17,* 65–70.

Hermans, H. J. M. (1976). *Prestatie motivatie test* [Performance motivation test]. Amsterdam: Swets & Zeitlinger.

Holding, D. H. (1985). *The psychology of chess.* Hillsdale, NJ: Erlbaum.

Holding, D. H., & Pfau, H. D. (1985). Thinking ahead in chess. *American Journal of Psychology, 98,* 271–282.

Howard, R. W. (1999). Preliminary real-world evidence that average human intelligence really is rising. *Intelligence, 27,* 235–250.

Ihaka, R., & Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics, 5,* 299–314.

Joireman, J. A., Fick, C. S., & Anderson, J. W. (2002). Sensation seeking and involvement in chess. *Personality and Individual Differences, 32,* 509–516.

Jonker, H. (1992). *Het Nederlandse Elo-rating boek: Elo-ratings en het KNSB-ratingssysteem* [The Dutch Elo rating book: Elo ratings and the KNSB rating system]. Venlo, The Netherlands: Uitgeverij van Spijk b.v.

Jöreskog, K. G., & Sorbom, D. (1996). *LISREL 8: User's reference guide.* Hillsdale, NJ: Erlbaum.

Karpov, A., & Mazukewitch, A. (1987). *Stellungsbeurteilung und Plan* [Assessment and plan]. Berlin: Sportverlag.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

McGregor, S. J., & Howes, A. (2002). The role of attack and defense semantics in skilled players' memory for chess positions. *Memory & Cognition, 30,* 707–717.

Meng, X., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin, 111,* 172–175.

Newell, A., & Simon, H. A. (1972). *Human problem solving.* Englewood Cliffs, NJ: Prentice Hall.

Nichelli, P., Grafman, J., Pietrini, P., Alway, D., Carton, J. C., & Miletich, R. (1994). Brain activity in chess playing. *Nature, 369,* 191.

Pfau, H. D., & Murphy, M. D. (1988). Role of verbal knowledge in chess skill. *American Journal of Psychology, 101,* 73–86.

Portisch, L, & Srközy, B. (1986). *Enspiele.* Thun, Switzerland: Harri Deutsch.

Robbins, T. W., Anderson, E. J., Barker, D. R., Bradley, A. C., Fearnyhough, C., Henson, R., Hudson, S. R., & Baddeley, A. D. (1996). Working memory in chess. *Memory & Cognition, 24,* 83–93.

Saariluoma, P. (1989). Chess players' recall of auditorially presented chess positions. *European Journal of Psychology, 1,* 309–320.

Saariluoma, P. (1995). *Chess players' thinking.* London: Routledge.

Saariluoma, P., & Kalakoski, V. (1998). Apperception and imagery in blindfold chess. *Memory, 6,* 67–90.

Simon, H. A., & Chase, W. G. (1973). Skill in chess. *American Psychologist, 61,* 394–403.

Sonas, F. (2002). The Sonas rating formula: Better than Elo? *Chessbase News*. Retrieved from http://www.chessbase.com/newsdetail.asp?newsid=562

Suetin, A. (1976). *Schachstrategie für Fortgeschrittene* [Chess strategy for advanced players (Vols. 1 and 2). Berlin: Sportverlag.

Thurstone, L. L. (1994). A law of comparative judgment. *Psychological Review, 101,* 266–270. (Original work published 1927)

Verstralen, H. H. F. M., Verhelst, N. D., & Bechger, T. M. (2001). *A double hazard model for mental speed.* Arnhem, the Netherlands: CITO Measurement and Research Department Reports.

Vicente, K. J., & de Groot, A. D. (1990). The memory recall paradigm: Straightening out the historical record. *American Psychologist, 45,* 285–287.

Vicente, K. J., & Wang, J. H. (1998). An ecological theory of expertise effects in memory recall. *Psychological Review, 105,* 33–57.

Wainer, H. (1990). *Computer adaptive testing: A primer.* Hillsdale, NJ: Erlbaum.

Wickelgren, W. A. (1977). Speed–accuracy tradeoff and information processing dynamics. *Acta Psychologica, 41,* 67–85.