Does Experience Teach? Professionals and Minimax Play in the Lab*

John Wooders†

June 6, 2008

Abstract

Does expertise in strategic behavior obtained in the field transfer to the abstract setting of the laboratory? Palacios-Huerta and Volij (2008) argue that the behavior of professional soccer players in mixed-strategy games conforms closely to minimax play while the behavior of students (who are presumably novices in strategic situations requiring unpredictability) does not. We reexamine their data, showing that, in fact, the play of professionals is inconsistent with the minimax hypothesis in several important respects: (i) professionals follow non-stationary mixtures, with card frequencies that are negatively correlated between the first and the second half of the experiment; (ii) professionals tend to switch between halves from underplaying a card relative to its equilibrium frequency to overplaying it (and vice-versa), and (iii) the distribution of card frequencies across professionals is far from the distribution implied by minimax. In each of these respects the behavior of students conforms more closely to the minimax hypothesis.

*I've benefited from many discussions over the years with Mark Walker about mixed-strategy play, and my thoughts on this topic have been influenced by those discussions. I'm grateful to Diego Moreno, Matt Van Essen, and Mark Walker for helpful comments and to Matt for outstanding research assistance.

†Department of Economics, Eller College of Business & Public Administration, University of Arizona, Tucson, AZ 85721 (jwooders@eller.arizona.edu).

1 Introduction

Several recent papers have established that the behavior on the field of sports professionals in strategic situations requiring unpredictability is consistent with the minimax hypothesis and its generalization to the theory of mixed-strategy Nash equilibrium. (See Walker-Wooders (2001) and Hsu-Huang-Tang (2007) for tennis, and Chiapppori, Levitt, and Groseclose (2002) and Palacios-Huerta (2003) for soccer). This raises an important question: Does expertise in strategic behavior obtained in a familiar setting, e.g., the tennis court or the soccer field, transfer to an unfamiliar one? If it does, then a key implication is that the nature of the subject pool is a critical ingredient of whether results obtained in the laboratory are useful for predicting behavior in the field.

In an ingenious experiment, Palacios-Huerta and Volij (2008) recruited professional soccer players and students to play two mixed-strategy games in the lab, and they obtained extraordinary results. The data shows, so they argue, that the play of professionals conforms remarkably closely to the behavior predicted by the theory whereas the play of student subjects does not. This suggests that the expertise in mixed-strategy play developed (and demonstrated) by professionals on the field does transfer from the field to the abstract setting of the laboratory.

The present paper re-examines the Palacios-Huerta and Volij (henceforth PH-V) data and argues that the behavior of soccer professionals is inconsistent with the minimax hypothesis in several important respects: (i) professional follow non-stationary mixtures, with card frequencies that are negatively correlated between the first and the second half of the experiment; (ii) they tend to switch between halves from underplaying a card relative to its equilibrium frequency to overplaying it (and vice-versa), and (iii) the distribution of card frequencies across players is far from the distribution implied by minimax. In each of these respects the behavior of students conforms more closely to the minimax hypothesis than does the behavior of professionals.

Perhaps paradoxically, this re-examination is motivated by the fact that actual play is *too close* to the theoretically expected play for soccer profession-

als. To illustrate this idea, consider the behavior of the soccer professionals in the PH-V data when playing the O'Neill game. In this game, to be described shortly, the minimax hypothesis calls for each player to choose cards (*1*, *2*, *3*, or *Joker*) according to an *iid* mixture which assigns probability .4 to the *Joker* card, and probability .2 to each of the *non-Joker* cards. In PH-V's experiment 40 professionals, in 20 pairs, played the O'Neill game 200 times and so 80 *Jokers* are expected for each player. However, the probability that a player chooses almost exactly 80 *Jokers*, say between 79 and 81 Jokers, is only 0.171 and so we would expect only 6.85 players for whom actual play was this close to expected play. Surprisingly, in the PH-V data 16 professionals choose between 79 and 81 *Joker*s. Such an outcome is extremely unlikely under minimax play.[1] The professionals' empirical card frequencies exhibit the same tendency to be too close to the theoretically expected frequencies for each of the three *non-Joker* cards as well.

PH-V suggest that professionals do not literally follow the *iid* minimax mixture, but rather they "...try to 'match' some probabilities they have, consciously or unconsciously, in mind." In other words, professionals choose cards as though a law of small numbers applies, keeping their empirical card choices close to the expected frequencies.

To investigate this hypothesis we partition the data into two halves, consisting of the first and last 100 rounds, respectively. If professionals indeed try to match the equilibrium frequencies, then play would be expected to conform closely to equilibrium in each half. Moreover, the striking difference found between professionals and students in the overall data should also hold in each half. If, perhaps, the adjustment by professionals to equilibrium play is time consuming, then we would at least expect that play conforms closely to equilibrium for the last 100 rounds.

We find no evidence that professionals match minimax frequencies when either half of the data is considered in isolation. Even though tests based on only half the data have less power, the hypothesis that professionals choose each card individually (or all four cards jointly) according to its minimax

[1]The probability that 16 or more players choose between 79 and 81 Jokers is only 0.00053, i.e., about 1 in 1900.

frequency is rejected at a high rate in each half of the data. Indeed, in each half of the data these hypotheses are rejected at similar rates for professional and students, and at too high a rate to be consistent with the minimax hypothesis.

Considering all 200 rounds, PH-V show that the joint hypothesis that all 40 professionals choose their cards in the O'Neill game according to the minimax model is not rejected (the p-value is .988) when the Pearson goodness of fit test is applied to their empirical card frequencies. In other words, if a computer were programmed to choose cards according to the true minimax frequencies, the odds of generating data in which the empirical card frequencies were this close (or closer) to the theoretically expected frequencies is only 1.2%. The same hypothesis is rejected at the 1% level for students (p-value 0.006). How is such a striking difference between professionals and students possible when the card choices of each exhibit a similar conformity to equilibrium in each half of the data? The answer is found by seeing that the behavior of professionals and students does differ in important ways.

The Minimax Hypothesis, applied to the repeated O'Neill game, requires that each player choose cards according to the same *iid* mixture at each round. An implication is that the frequency with which a card is chosen in the first half is uncorrelated with the frequency it's chosen in the second half. This hypothesis is rejected for professionals, for each of the four cards in the O'Neill game, in favor of negative correlation. For students, in contrast, this hypothesis is only rejected for the *Joker* card (for which there is a positive correlation between the first and second half frequencies). Thus, with respect to whether the first and second half card frequencies are uncorrelated, the behavior of students conforms more closely to the theory than does the behavior of professionals.

A second and related implication is that if a subject happens by chance to play a card with, say, less than its equilibrium frequency in the first half (i.e., he "underplays" the card), this has no bearing on the likelihood he will underplay it in the second half. We show that professionals vary their play systematically over the two halves of the experiment. Professionals who underplay the *Joker* card relative to its equilibrium frequency in the first

100 rounds, tend to overplay it in the last 100 rounds (and vice versa). As a result, the frequency with which professionals choose the *Joker* card over all 200 plays is closer to the equilibrium frequency than it is in either half alone. For each of the *non-Joker* cards, professionals also exhibit the same tendency to switch between underplaying and overplaying it. Students, in conformity with the theory, exhibit no systematic tendency to switch.

Considering either half of the data alone, the play of both groups is similar in the degree to which it conforms to equilibrium. The essential difference between professionals and students is how their play changes between the first half and second half. For students we observe no change, while the tendency of professionals to change their play across halves has a powerful effect – it causes their overall card frequencies to be excessively closely clustered around the expected card frequencies. Applying the Kolmogorov-Smirnov goodness of fit test to the overall card choices of professionals, one can reject minimax play for each card individually and for all four cards jointly. The same test applied to the card choices of students yields a rejection only for the *Joker* card.

In Section 2 we describe the PH-V experiment. In section 3 we compare the behavior of professionals and students in the first and last 100 rounds, showing that the behavior of each conforms equally well to the theory in each half. In Section 4 we show that the empirical card frequencies of professionals are negatively correlated between halves, professionals switch between underplaying and overplaying a card relative to its equilibrium frequency, and the distribution of card frequencies is inconsistent with minimax play. Section 5 considers the data for PH-V's "Penalty Kick" game and concludes.

2 The PH-V Experiment

PH-V recruited professional soccer players and college students to play two zero-sum games, a "Penalty Kick" game which they introduce, and the well-known O'Neill game (O'Neill 1987). The Penalty Kick game is a stylized representation of a penalty kick in soccer where the kicker (Row) chooses

whether to kick left (A) or right (B), and the goalie (Column) simultaneously chooses whether to cover left (A) or right (B). The O'Neill game is a zero-sum game with a unique asymmetric mixed-strategy Nash equilibrium.

	A	*B*
A	.60	.95
B	.90	.70

Penalty Kick Game

	Red	*Brown*	*Purple*	*Green*
Red	0	1	1	0
Brown	1	0	1	0
Purple	1	1	0	0
Green	0	0	0	1

The O'Neill Game

In both games of the PH-V experiment the payoff numbers are the probability that the row player wins 1 Euro.

Eighty professional soccer players were recruited, 40 of whom were kickers and 40 of whom were goalies. For the Penalty Kick game, 40 professionals in 20 fixed pairs, with a kicker in the row role and a goalie in the column row played 150 rounds. For the O'Neill game, another 40 professionals were paired in the same fashion but played 200 rather than 150 rounds. A total of 160 college students participated, half with soccer experience and half without. To sharpen the contrast, we focus on the 80 students without soccer experience. Like the professionals, 20 fixed pairs of students played 150 rounds of the Penalty Kick game and 20 fixed pairs played 200 rounds of the O'Neill game. In both games the subjects played 15 practice rounds. Subjects were not told the number of rounds to be played.

The PH-V experiment improves on O'Neill's (1987) original study in several respects. First, the stakes are much higher – in O'Neill (1987) subjects played for 5 cents per game. Second, in order to avoid the *Ace* bias noted in Brown and Rosenthal (1990), PH-V follow Shachat (2002) in labeling the strategies *Red, Brown, Purple, and Green*, rather than *Ace*, *Two*, *Three*, and *Joker*.[2] For expositional convenience, we follow PH-V and use O'Neill's

[2]See Brown and Rosenthal (1990) for a re-examination of O'Neill's experiment and see O'Neill (1991) for a response.

original labelling with *1* for *Ace*, *2* for *Two*, *3* for *Three*, and *J* for *Joker*.[3]

An innovation of O'Neill's pioneering study was that the stage game had only two outcomes, i.e., each player either won or lost, and hence the minimax (and Nash equilibrium) mixture did not depend on the players' attitude toward risk. The supergame consisting of the repeated play of the stage game clearly has more than two outcomes and hence one might wonder whether the players' risk attitudes once again become important. Wooders and Shachat (2001) shows if that the stage game has two outcomes, is strictly competitive (i.e., if one player wins then the other loses), and has a unique Nash equilibrium, then the unique Nash equilibrium of the supergame calls for Nash play at each stage provided the players' preferences over supergame outcomes satisfy a weak monotonicity condition.[4] The unique Nash equilibrium of the 200-time repeated O'Neill game, for example, is for each player to choose the *Joker* card at each stage with probability .4 and choose each of the non-*Joker* cards with probability .2, independently of the history of play up to that point.

3 Comparing Professionals and Students

PH-V argue that the behavior of professionals in the O'Neill game conforms closely with theory, both at the level of an individual player and in aggregate, while the behavior of students is far from equilibrium. Considering all 200 rounds, they find that the null hypothesis that a subject chooses the *Joker* card with probability .4 is rejected at the 5% level for 4 professionals, but the same null is rejected for 9 students. For the *non-Joker* cards, the minimax

[3]Other important experimental studies of mixed strategy play include Rapoport and Boebel (1992), Mookherjee and Sopher (1994), Ochs (1994), Rosenthal, Shachat, and Walker (2003), and Shachat (2002). See Camerer (2003) for an in-depth survey.

[4]The monotonicity condition is easily illustrated in the twice-repeated O'Neill game, which has four outcomes: A player can win twice ($\mathcal{WW}$), he can win and then lose ($\mathcal{WL}$), or lose and then win ($\mathcal{LW}$), or lose twice ($\mathcal{LL}$). Monotonicity requires $\mathcal{WW} \succ \mathcal{WL} \succ \mathcal{LL}$ and $\mathcal{WW} \succ \mathcal{LW} \succ \mathcal{LL}$. Monotonicity allows players to be risk averse, preferring the outcome $\mathcal{WL}$ to a 50-50 lottery on $\mathcal{WW}$ and $\mathcal{LL}$, and allows players to be impatient, preferring $\mathcal{WL}$ to $\mathcal{LW}$.

binomial model is rejected at the 5% level in 4 instances for professionals, while it is rejected in 14 instances for students.[5] Applying the Pearson Goodness of fit test to each subject's choice of all four cards, the null hypothesis that the subject choose all four cards according to the equilibrium mixture is rejected at the 5% level for only 2 professionals, while it is rejected for 7 students. The joint hypothesis that all 40 players choose their cards according to the equilibrium mixture is not rejected for professionals (p-value 0.988), while it is rejected at the 1% level for students (p-value 0.006).[6]

As noted earlier, a striking feature of the PH-V data is that the actual play of professionals is "too close" to expected play to be consistent with the players following the minimax *iid* mixture. There were 16 professionals who chose between 79 and 81 *Jokers*, whereas we would expect only 6.85 under minimax play. The probability that 16 or more (of 40) subjects play within one card of the expected number of *Jokers* is only 0.00053. For the *non-Joker* cards *1*, *2*, and *3*, there were, respectively, 21, 16, and 18 professionals who chose within one card of the expected number. The probability that 16 or more subjects play within one of the expected number of *2* cards is only 0.004642 (only 8.35 such players are expected), and the analogous probabilities for the *1* and *3* cards are much lower. Under minimax play it is extremely unlikely that for any one of the cards there would be so many players for whom actual play is in such close correspondence with expected play. The close correspondence of actual to expected play is especially striking given that it holds for all four cards.

PH-V propose that "The excessive closeness of the observed frequencies to the hypothesized ones suggests that subjects do not randomize, but rather try to 'match' some probabilities they have, consciously or unconsciously, in mind." If professionals were matching frequencies, then one might expect to find serial correlation their choices but, employing a runs test, PH-V do not reject randomness. In this section we examine whether the PH-V data is consistent with professionals choosing cards to "match" the minimax frequencies.

[5] These results are reported in Table X and XIV, respectively, in PH-V.

[6] See Table 3 of the present paper for these p-values.

An important aspect of PH-V's experimental design was that subjects were *not* told the number of hands they would play. To see the significance of this feature, consider a repeated matching pennies game in which the subjects are told they will play 200 times. A subject aiming to match the minimax frequencies could do so by producing 100 Heads over the 200 rounds. The subject, however, might not match the minimax frequencies over any particular sequence of rounds – he might choose substantially more than 50 Heads over the first 100 rounds, say. If the number of rounds is not known, then a subject aiming to match the minimax frequencies would need to choose faces so that roughly half of his choices were Heads at any given point in the experiment. In the O'Neill game, analogously, if a subject is matching the minimax frequencies then we should find over any long interval of play that roughly 40% of the cards are *Joker*s and that the frequencies of the other cards are close to their minimax frequencies as well.[7]

Here we partition the data into the first 100 rounds (i.e., the first half of the experiment) and the last 100 rounds. If professionals are indeed matching frequencies, then the close conformity found over the whole experiment of their empirical card frequencies to the expected frequencies should also be found when each half of the data is considered in isolation. The hypothesis that professionals choose an individual card (e.g., the *Joker* card) according to the minimax binomial model should be rejected at the same (or perhaps even a lower) rate in each half of the data than it was in the data overall.[8] Similarly, the striking overall difference between professionals and students in the degree to which their play conforms to minimax should also be found when each half of the data is considered in isolation.

A Randomized Binomial Test

We will be interested in testing whether the choice frequency of each individual card is consistent with minimax play. To do so it is useful to in-

[7] It would be very difficult to match the minimax frequencies, while at the same time remaining unpredictable to ones opponent, even if a subject knew the frequencies and that 200 rounds would be played.

[8] Since each half of the data has only 100 plays, all else equal our tests will have lower power and hence we would expect *fewer* rejections of the null hypothesis.

troduce a randomized binomial test. Under the null hypothesis of minimax play, the number of *Joker* cards chosen after 200 rounds is distributed Binomial $B(n,p)$, with *cdf* denoted by $F(n_J;n,p)$, where $n=200$, $p=.4$, and n_J is the number of Jokers. If n_J^i is the number of Jokers chosen by subject i, we form the random test statistic t^i where $t^i \sim U[0, F(0;200,.4)]$ if $n_J^i = 0$ and $t^i \sim U[F(n_J^i - 1;200,.4), F(n_J^i;200,.4)]$ otherwise. Under the null hypothesis of minimax play, prior to the realization of n_J^i the statistic t^i is distributed $U[0,1]$.[9] For each t^i, the associated p-value is $p^i = \min\{2t^i, 2(1-t^i)\}$, which is also distributed $U[0,1]$.[10] We reject the null hypothesis at significance level α if $p^i \leq \alpha$. This procedure generalizes in the obvious way to yield a randomized test for each of the *non-Joker* cards.

For nearly all realizations of n_J^i, this randomized binomial test and a deterministic decision rule of the same (approximate) size in which the null is rejected if there are either too many or too few *Jokers* leads to the same decision. Using the randomized test, if $\alpha = .05$ and $n_J^i = 66$ then $t^i \sim U[0.0173, 0.0247]$, and $p^i \sim U[0.0346, 0.0494]$, and the null is rejected for every realization of p^i. If $n_J^i = 67$ then $t^i \sim U[.0247, .0346]$, and $p^i \sim U[0.0494, 0.0692]$, and the null is rejected with probability .030369. In contrast, if we follow the deterministic rule of rejecting the null if there are fewer than 67 or more than 93 *Jokers* (a test of size .05105), we reject the null if $n_J^i = 66$ but do not reject it if $n_J^i = 67$. Only when $n_J^i = 67$ or $n_J^i = 94$ will the two

[9] Let $z \in [0,1]$ be arbitrary. We show that $\Pr[t^i \leq z] = z$. To simplify notation, write $F(n_J)$ for $F(n_J;200,.4)$. Let $k=0$ if $z \leq F(0)$ and let $k \in \{1,\ldots,200\}$ be such that $F(k-1) < z \leq F(k)$ otherwise. Observe that if $n_J^i < k$ then $\Pr[t^i \leq z | n_J^i] = 1$; if $n_J^i = k$ then $\Pr[t^i \leq z | n_J^i] = \frac{z-F(k-1)}{F(k)-F(k-1)}$; if $n_J^i > k$ then $\Pr[t^i \leq z | n_J^i] = 0$. Hence

$$\begin{aligned} \Pr[t^i \leq z] &= \sum_{j=0}^{k-1} \Pr[n_J^i = j] + \frac{z-F(k-1)}{F(k)-F(k-1)} \Pr[n_J^i = k] \\ &= F(k-1) + \frac{z-F(k-1)}{F(k)-F(k-1)}[F(k)-F(k-1)] \\ &= z. \end{aligned}$$

[10] To see this, note that $\Pr[p^i \leq z] = \Pr[\min\{2t^i, 2(1-t^i)\} \leq z] = \Pr[\min\{t^i, 1-t^i\} \leq \frac{z}{2}] = \Pr[t^i \leq \frac{z}{2}] + \Pr[t^i \geq 1 - \frac{z}{2}] = z$ since t^i is distributed $U[0,1]$.

decision rules lead to different decisions with positive probability.

A randomized binomial test has nonetheless two advantages over a deterministic decision rule. First, even with a finite sample, the randomized test is symmetric and of exactly size α.[11] More important, under the null that each player chooses *Joker* with probability .4, then each p^i is drawn from the same continuous distribution (*viz.* the $U[0,1]$ distribution) and hence we can apply the Kolmogorov-Smirnov (KS) goodness of fit test to determine whether the empirical *cdf* of the 40 values of p^i differs from the theoretical one.

The O'Neill Game – First half

Tables 1 and 2 show the card choice frequencies of professionals and students, respectively, in first and last 100 rounds of the O'Neill game. We begin by focusing on play in the first half, which is reported on the left hand side of these tables. The null hypothesis that a player chooses *Joker* according to the minimax binomial model in each of the first 100 rounds is rejected at the 5% level for 7 professionals and 6 students (in each case only 2 rejections are expected under minimax play). For the *1*, *2*, and *3* cards, the null hypothesis that each of these cards is chosen according to the minimax binomial model is rejected at the 5% level in 6 instances for professionals and 4 instances for students. Under the null, a total of 6 rejections, two rejections per *non-Joker* card, are expected at the 5% level.

Next we consider the joint hypothesis that all 20 professionals in a given role choose a given card according to the minimax binomial model.[12] Since there are two roles and four cards, 8 different null hypotheses are considered.

[11] PH-V reject the null that the *Joker* card is chosen according to the minimax binomial model at the 5% significance level if a player chooses fewer than 68 or more than 93 *Jokers*. The test is not symmetric since the probabilities of these two events, 3.4594×10^{-2} and 2.4716×10^{-2}, respectively, are unequal. Moreover, the size of the test is .06 rather than .05. The test would have been closer to the correct size if the rule had been to reject the null whenever there were fewer than 67 *Jokers*.

[12] Let n_J^i denote the number of *Joker* cards, say, played by professional i in the first 100 rounds. Under the null $n_J^i \sim B(100, .4)$ for each i, and hence the joint null is that $\sum_{i=1}^{20} n_J^i \sim B(2000, .4)$. We report the realized random p-values.

This null is rejected at the 5% level for column players choosing the *Joker* card (p-value 0.00017), the *1* card (p-value 0.030), the *2* card (p-value 0.020), and for row players choosing the *3* card (p-value 0.014). Pooling the choices of the two roles, the joint null that all 40 professionals choose a given card according to the minimax binomial model is rejected only for the *2* card. For students, in contrast, the same hypotheses are not rejected for any for the four cards, in either role, or when the choices of the two roles are pooled.

Pearson goodness of fit tests of the null hypothesis that a player chooses all four cards in the first 100 rounds according to the minimax multinomial model are reported in Table 3. (The table reports the p-values for these tests.) Let n_1^i, n_2^i, n_3^i and n_J^i denote the number of times in the first 100 rounds that subject i chose the *1*, *2*, *3*, and *Joker* cards, respectively, and let p_1^i, p_2^i, p_3^i and p_J^i denote the true (but unknown) probability with which subject i chooses each card. Under the null hypothesis that player i chose cards according to the minimax mixture, i.e., $p_1^i = p_2^i = p_3^i = .2$ and $p_J^i = .4$, the statistic

$$Q^i = \sum_{s \in \{1,2,3,J\}} \frac{(n_s^i - 100p_s^i)^2}{100p_s^i},$$

is asymptotically distributed chi-square with three degrees of freedom.

For the first 100 plays this null is rejected at the 5% level for 3 professionals and 3 students. The joint null hypothesis that all 20 players in a given role choose all four cards according to the minimax multinomial model is rejected at the 5% level for the row role, for both professionals and students. The analogous null is not rejected for the column role or when the roles are combined, for either professionals or students.

These results show that during the first 100 plays there is little difference between professionals and students in terms of the number of 5% rejections of either the binomial minimax model for single cards, or the multinomial minimax model for all four cards jointly. For both groups we obtain more than the expected number of rejections of the minimax binomial model for the *Joker* card. If anything, aggregate play is closer to equilibrium for students – the joint null hypothesis that all the students in a given role choose a given card according to the minimax binomial model is not rejected for any of the

cards, for either the column or the row role (or when the roles are pooled), while the same null is often rejected for professionals.

The O'Neill Game – Second half

Perhaps professionals have learned to play equilibrium by the second half of the experiment, while students have not. To assess this possibility, we apply the same statistical tests to the last 100 rounds. (Play in the second half is given on Tables 1 and 2 under the "Second Half" heading.) The null hypothesis that players choose the *Joker* card according to the minimax binomial model is rejected at the 5% level for 7 professionals and 5 students. For the *non-Joker* cards, the minimax binomial model is rejected in 3 instances for professionals and in 6 instances for students. This is similar to the number of rejections we saw in the first half (6 for professionals and 4 for students). We now see slightly more rejections for students, although the number of rejections does not exceed the number expected under the null.

Next we consider the joint null hypothesis that all the players in a given role choose a given card according to the minimax binomial model. We obtain 4 rejections (of 8 possible) at the 5% level for professionals – we reject for column playing the *1*, *2*, and *Joker* cards and for row playing the *Joker* card. We obtain only 2 rejections for students – column playing the *3* card and the *Joker* card. However, pooling the choices of both roles, we obtain no rejections for professionals while we obtain two rejections (for the *1* card and the *Joker* card) for students.

Table 3 shows that the null hypothesis that all four cards are chosen according to the minimax multinomial model is rejected for 3 professionals and for 4 students (2 rejections are expected). For both professionals and students, the joint null hypothesis that all players in a given role choose cards according to the minimax multinomial model is not rejected at the 5% level for either the column role, the row role, or when the roles are combined.[13]

These results show that in the last 100 rounds the behavior of professionals and students, at the individual level, is very similar in terms of the

[13]The p-value for students in the Row role is .051 and hence only just barely does the null fail to be rejected.

number of 5% rejections of the minimax binomial model for individual cards and the minimax multinomial model for all four cards. For both groups there are more rejections than expected of the minimax binomial model for the *Joker* card.

The table below summarizes the results for the joint tests of the minimax binomial and multinomial models, for the first and second half of the data. A rejection at the 5% significance level is indicated by a "×" mark, where the column headings "R," "C," and "P" stand for row, column, and Pooled, respectively.

	Professionals						Students					
	1^{st} Half			2^{nd} Half			1^{st} Half			2^{nd} Half		
Card	R	C	P	R	C	P	R	C	P	R	C	P
1		×			×							×
2		×	×									
3	×				×						×	
J		×		×	×						×	×
1-2-3-J	×						×					

Table 4: 5% Rejections of the Joint Minimax Binomial/Multinomial Model

In both halves, and especially for professionals in the first half, these joint tests are rejected more frequently than we would expect under the theory.

KS Tests on the First and Last 100 Plays

An alternative test of the joint hypothesis that all the professionals choose a given card according to its minimax probability is based on the empirical distribution of the 40 values of p^i obtained from applying the randomized binomial test to each professional's play. Under the null hypothesis that professional i chooses the *Joker* card, say, according to the minimax binomial model, then the p-value p^i obtained from the randomized binomial test is distributed $U[0,1]$. The Kolmogorov-Smirnov (KS) test allows us to test whether the empirical distribution of the p^i's is generated according to the theory – i.e., the uniform distribution whose *cdf* is given by the 45° line.

Formally, the KS test is as follows: The hypothesized *cdf* for the p-values is the uniform distribution, $F(x) = x$ for $x \in [0, 1]$. The randomized binomial test for a given card yields 40 p-values, one for each player. The empirical distribution of these p-values, denoted $\hat{F}(x)$, is given by $\hat{F}(x) = \frac{1}{40}\sum_{i=1}^{40} I_{[0,x]}(p^i)$, where $I_{[0,x]}(p^i) = 1$ if $p^i \leq x$ and $I_{[0,x]}(p^i) = 0$ otherwise. Under the null hypothesis, the test statistic $K = \sqrt{40}\sup_{x\in[0,1]} |\hat{F}(x) - x|$ has a known distribution (see p. 509 of Mood, Boes, and Graybill (1974)).

In addition to its visual appeal, the KS joint test has several advantages over joint tests based on the binomial distribution (for a single card) or the Pearson goodness of fit test (for all four cards jointly) used above. First, the minimax hypothesis generates a prediction about the *distribution* of card frequencies across players, and the KS test can be applied to determine whether the empirical distribution of card frequencies matches the predicted one. The Pearson joint test, in contrast, focuses on only one aspect of the distribution – its mean. To illustrate this point, suppose that every subject in the PH-V experiment chose *exactly* 80 Joker cards after 200 plays. The joint null hypothesis that each player chose the *Joker* card according to the minimax binomial model would not be rejected by the Pearson joint test, even though such an outcome is clearly inconsistent with each of the players following an *iid* mixture with probability .4 on the *Joker* card. The same joint null would, however, be rejected by the KS test since the empirical distribution of the p-values (i.e., the *cdf* which assigns all probability to p-values of 1) is not close to the uniform distribution.

Second, the KS test is not sensitive to outliers since the empirical *cdf* of p-values is little changed by the addition or removal of single p-value. In contrast, since the Pearson joint test is based on the sum of the test statistics of the individual players, the value of the statistic is sensitive to outliers. For example, the minimax multinomial model is rejected for students on the basis of the overall data (p-value of 0.006), but if we exclude the row player in pair 10 the p-value becomes .07 and the null is no longer rejected. The Binomial joint test is based on the total number of times a given card is played, and so it is also sensitive to the effects of outliers.

The top panel of Figure 1 shows the empirical *cdf* of the p's for the

randomized binomial test applied to the *Joker* choices of professionals and students, for the first 100 rounds. The bottom panel shows the empirical *cdf* of p-values (40 for professionals and 40 for students) obtained by applying the Pearson goodness of fit test to the choices of all four cards jointly. Figure 2 shows the empirical *cdf* of the p's when the randomized binomial test is applied to each of the non-*Joker* cards, for both professionals and students.

Table 5 reports the results of the KS test applied to each of the four cards individually, as well as all four cards jointly. For professionals, we cannot reject the joint null hypothesis that cards are chosen according to the minimax binomial model for any of the cards (or all cards jointly). For students, minimax play is rejected for the *Joker* card and for the *2* card. For the *Joker* card there are too many small p-values; we have $\hat{F}(.21) = .475$, i.e., 47.5% of the p-values (19 of the 40 values) are less than or equal to .21, whereas only 8.4 such values are expected. The KS test shows that this is a statistically significant difference. For the *2* card there are too many large p-values; we have $\hat{F}(.68) = .43$, i.e., 57% of the p-values (23 of 40 values) are greater than or equal to .68. In particular, actual play was too close to expected play, with 22 students choosing the card either 39, 40, or 41 times.

	Professionals			Students	
	KS	p-value		KS	p-value
J	1.153	0.140321	*J*	1.690	0.006625
1	1.279	0.075907	*1*	1.165	1.164581
2	0.891	0.404775	*2*	1.623	0.010276
3	0.832	0.493269	*3*	0.852	0.462152
1-2-3-J	1.012	0.256921	*1-2-3-J*	0.734	0.654362

Table 5: First Half – KS tests of conformity to $U[0, 1]$

In contrast to the binomial and multinomial joint tests presented earlier, the KS test identifies a dimension in which the play of professionals is closer to equilibrium than the play of students in the first 100 rounds.

Figures 3 and 4 show the same *cdf*'s for professionals and students for the last 100 rounds. A visual comparison shows a remarkably similarity between

the *cdf*'s of the two groups for all four cards, both individually and jointly. And, with the exception of the *Joker* card, the empirical *cdf*'s closely follow the theoretical *cdf*. Table 6 shows that the joint null that each professional chooses the *Joker* card according to the minimax binomial model just fails to be rejected at the 5% significance level. The analogous joint null is not rejected for professionals or students for any other of the four cards. The joint null that all four cards are chosen according to the minimax multinomial model is also not rejected for either professionals or students.

	Professionals			Students	
	KS	p-value		KS	p-value
J	1.335	0.056760	*J*	1.090	0.185479
1	0.553	0.920282	*1*	1.124	0.159444
2	0.707	0.699818	*2*	0.633	0.817785
3	0.689	0.729337	*3*	0.680	0.743792
1-2-3-J	0.964	0.311085	*1-2-3-J*	0.930	0.353073

Table 6: Second Half – KS tests of conformity to $U[0,1]$

The empirical *cdf*s in Figures 1 through 4 show that professionals are not, as PH-V suggest, choosing cards to "match" the minimax frequencies in either the first or the second half of the experiment. If they were frequency matching, the *cdf*'s would be characterized by too many large p-values, i.e., the empirical *cdf* $\hat{F}(x)$ would lie far below the theoretical *cdf* $F(x)$. In fact, a visual inspection reveals that the empirical and theoretical *cdf*'s are generally close, for both the first and second half, and for all four cards individually and jointly.

4 Resolving the Puzzle

We have seen that the empirical card frequencies of professionals and students exhibit a similar conformity to equilibrium, both in the first 100 rounds and in the last 100 rounds. Yet, PH-V have shown that when the binomial or multinomial Pearson goodness of fit tests are applied to the overall O'Neill

data, the behavior of professionals appears to conform closely to equilibrium while the behavior of students does not. How is this possible? Moreover, if professionals are not matching frequencies, then why are their empirical card frequencies too close to the theoretically expected frequencies?

In fact, the behavior of professionals and students *does* differ in important ways. Moreover, it differs in a way that explains both why there is a dramatic difference between students and professionals in the data overall and why the empirical card frequencies of professionals are "too close" to the expected frequencies.

In the equilibrium of the repeated O'Neill game, each player chooses cards according the same *iid* mixture at each stage. An implication is that the empirical frequency with which a player chooses the *Joker* card, say, in the first half is uncorrelated with the frequency with which he chooses it in the second half. The top panel of Figure 5 shows these frequencies for professionals, with each point (x, y) representing the first- and second-half *Joker* frequencies, x and y, respectively, of a professional. A point where $x < .4$ and $y > .4$ falls in the upper-left quadrant and corresponds to a player who underplayed *Joker* in the first half, but overplayed *Joker* in the second half. A point on a line with slope -1 that passes though $(.4, .4)$ corresponds to a player whose choice frequencies differ from minimax in each half, but which matches the minimax frequency overall.

Figure 6 plots the same frequencies for the *non-Joker* cards, where the four quadrants are now defined relative to the .2 equilibrium mixture. For convenience, each figure also shows the linear regression line.

Table 7 below reports, for each card, the value of the Spearman rank correlation coefficient R between the frequency the card is played at the first half and the frequency it is played in the last half.[14] Under the null hypothesis that the first and second half frequencies are independent, the distribution of R is known and hence R can be used to obtain a non-parametric test of

[14] The calculation of the Spearman R corrects for ties in ranks, and is computed using the webpage http://faculty.vassar.edu/lowry/corr_rank.html, authored by Richard Lowry.

the null of independence. (See Gibbons and Chakraborti pp. 422-431.)

	Professionals				Students		
	R	t	p-value		R	t	p-value
J	-0.3195	-2.08	0.04432	*J*	0.3159	2.05	0.04731
1	-0.5804	-4.39	0.00009	*1*	0.2882	1.86	0.07064
2	-0.3688	-2.45	0.01900	*2*	-0.0160	-0.10	0.92087
3	-0.3463	-2.28	0.02831	*3*	-0.0427	-0.26	0.79627

Table 7: Spearman Rank Correlation Coefficients

The choice frequencies of professionals are negatively correlated between halves for each of the four cards, with the null hypothesis of no correlation rejected for each card at the 5% (or smaller) significance level. For students, the same null hypothesis is not rejected for any of the *non-Joker* cards. For the *Joker* card, however, the null hypothesis of no correlation is rejected at the 5% significance level, with the first-half and second-half frequencies positively correlated.

A second implication of the minimax hypothesis is that if a subject plays a card with, say, a frequency below its equilibrium frequency in the first half (i.e., he underplays it), this has no bearing on the likelihood he will underplay the card in the second half. In other words, a player's choice frequencies are equally likely to fall in each one of the four quadrants. It is immediately visually apparent from Figures 5 and 6 that professionals who underplay a card in the first half relative to its equilibrium frequency tend to overplay it in the second half, and vice versa.

Table 8 shows the number of frequencies falling into each of the four quadrants. There were 14 professionals who underplayed *Joker* in the first half but overplayed it in the second half; there were 11 who switched from overplaying to underplaying *Joker*. (Entries where the players "switch" are shown in bold.) Hence 25 players switched from underplaying to overplaying *Joker* (or vice versa). Ignoring the four players whose choices frequencies fall on the boundary (i.e., which satisfy $x = .4$ or $y = .4$), we would expect only 18 of the remaining 36 players to make such a switch. The null hypothesis that

the *Joker* frequencies are uniformly distributed over the four quadrants just fails to be rejected at the 10% level for professionals (p-value 0.1116) using the Pearson goodness of fit test.[15] The same null hypothesis is decisively rejected for each of the *non-Joker* cards, with p-values for the *1*, *2*, and *3* card of 0.0008, 0.0002, and 0.0272, respectively.[16]

1^{st} Half/ 2^{nd} Half	Under/ Over	Under/ Under	Over/ Over	Over/ Under	Total	Q	p-value
J	**14**	6	5	**11**	36	6.0000	0.1116
1	**9**	2	4	**17**	32	16.7500	0.0008
2	**7**	3	2	**17**	29	19.4138	0.0002
3	**16**	5	5	**11**	37	9.1622	0.0272
	46	16	16	**56**	134		

Table 8: Professionals, Counts by Quadrant

As shown in Table 9, the distribution of the first and second half choice frequencies of students is, in contrast, far more consistent with minimax play. The null hypothesis that the frequencies are uniformly distributed over quadrants is not rejected for any of the *non-Joker* cards. The null is rejected for the *Joker* card since students tend to underplay *Joker* in the second half,

[15] One can reject the null hypothesis that it is equally like that a player switches as not; the probability of 25 or more switches in 36 trials is only 0.014 under the null.

[16] Although each test is meaningful on its own, the tests are not independent. A subject who switches from underplaying to overplaying *Joker* must necessarily switch from overplaying to underplaying at least one of the *non-Joker* cards.

irrespective of whether they under or overplayed it in the first half.

1^{st} Half/ 2^{nd} Half	Under/ Over	Under/ Under	Over/ Over	Over/ Under	Total	Q	p-value
J	**3**	16	7	**10**	36	10.0000	0.0186
1	**11**	8	9	**6**	34	1.5294	0.6755
2	**11**	5	10	**5**	31	3.9677	0.2650
3	**11**	9	6	**5**	31	2.9355	0.4017
	36	32	38	**26**	132		

Table 9: Students, Counts by Quadrant

Switching and the Consequences for Overall Play

When a subject overplays a card (relative to its equilibrium frequency) in the first half, but underplays it in the second half, then the overall frequency with which the card is played will tend to be closer to equilibrium than in either half alone. Moreover, the sample variance will tend to be too small relative to the sample variance under equilibrium play. Consider, for example, the 200-time repeated matching pennies game. Under equilibrium play, the expected number of heads is 100 and the variance is 50. If, instead, a subject chooses H with probability $.5+\gamma$ in the first 100 plays, with $0 \leq \gamma \leq .5$, but chooses H with probability $.5-\gamma$ in the last 100 plays, the expected number of heads remains 100, but the variance is reduced to $50(1-4\gamma^2)$. Thus, given a collection of subjects whose play varies in this fashion, there will tend to be too many subjects with approximately 100 heads after 200 plays. Under the null hypothesis of a fair coin, applying the Binomial test there will tend to be too many subjects with large p-values.[17]

Figures 7 and 8 show the empirical *cdf* of p-values for each card alone and for all four cards jointly. For professionals, except for the *Joker* card, the

[17] There will as well tend to be too few rejections of the null hypothesis. Suppose our decision rule is to reject the null if there are 86 or fewer or 113 of more heads after 200 plays. A simple calculation shows that if the null is true, it is rejected with probability 0.056. If, instead, subjects choose H with probability .8 in the first 100 plays and choose H with probability .2 in the last 100 plays, then the null is rejected with probability 0.017. In other words, against this alternative, the null is rejected with probability less than .056.

empirical *cdf* of p-values lies far below the theoretical *cdf*, which indicates the presence of too many large p-values. In fact, the empirical *cdf*'s of the p-values is closer to the 45 degree line for students than professionals for each of the four cards, with the exception of the *Joker* card where the distances are virtually the same.

Table 10 shows the results of applying the KS test to the distribution of p-values obtained when the randomized binomial test is applied to each card individually and the Pearson goodness of fit test is applied all four cards jointly. The KS test for the *Joker* card yields the same results for professional and students – in each case the null that all subjects choose *Joker* according to the minimax mixture is rejected. For professionals, the maximal distance between the empirical *cdf* and the uniform *cdf* is at an "x" value of .833 where the value of the *cdf* is .6. In other words, 40% of professionals have p-values above .833 whereas we would expect 16.7% to have such p-values. Since the maximum distance occurs where the empirical *cdf* lies below the *cdf* of the uniform distribution, the rejection is a result of there being too many high p-values. For students, in contrast, the maximal difference between the two *cdf*s is where $x = 0.365$ and the empirical *cdf* lies above the uniform, with too many low p-values.

	Professionals			Students	
	KS	p-value		KS	p-value
J	1.477	0.025429	*J*	1.484	0.024464
1	2.332	0.000038	*1*	1.239	0.092692
2	1.917	0.001285	*2*	1.160	0.135789
3	1.693	0.006456	*3*	1.071	0.201243
1-2-3-J	2.434	0.000014	*1-2-3-J*	1.292	0.071098

Table 10: Overall – KS tests of conformity to $U[0, 1]$

For each *non-Joker* card, the KS test resoundingly rejects the joint null hypothesis that professionals play the card according to the minimax binomial model. In each case, the null is rejected as a result of the empirical *cdf* having too many large p-values. The result for the *1* card is especially

dramatic, with 82% of the p-values above .544. (Recall that the negative correlation between first and second half play was greatest for the *1* card.) In contrast, the null that a card is chosen according to its equilibrium mixture is not rejected for any one of *non-Joker* cards for students. The hypothesis that all four cards are jointly chosen according to the equilibrium mixture is rejected for professionals (p-value of 0.000014), but it is not rejected for students at the 5% significance level.

In this section we have shown that the behavior of professional conforms less closely to minimax than does the behavior of students in several respects. First, the frequency with which professionals play a card in the first half is negatively (and statistically significantly) correlated with the frequency it is played in the second half. Second, professionals tend to switch from underplaying a card in the first half to overplaying it in the second half (or vice versa). Third, for each card, and for all cards jointly, we can reject that the distribution of p-values for professionals is uniform, as predicted by the theory. The behavior of students, in contrast conforms with equilibrium in all three respects, with the exception of the *Joker* card.

We conclude this section by noting that when subjects following non-stationary mixtures, an additional consequence will be reduction in the number of runs. Consider again the matching pennies game. If n_1 Heads and n_2 Tails are played in $n_1 + n_2$ rounds, then the expected number of runs under the null hypothesis of randomness is

$$\frac{2n_1n_2}{n_1 + n_2} + 1.$$

If, for example, $n_1 = 70$ and $n_2 = 30$ in the first 100 rounds, but the number of Heads and Tails is reversed in the last 100 rounds, then 43 runs are expected in each half and 86 runs are expected overall. If, instead, 100 Heads and 100 Tails are randomly distributed over 200 rounds, then 101 runs are expected. Hence when subjects follow non-stationary mixtures, this will introduce a bias towards too few runs (i.e., towards positive serial correlation) and away from the negative serial correlation commonly found in laboratory experiments with student subjects.

5 Discussion

The Penalty Kick Game

The differences between professionals and students found in the O'Neill game are exhibited in the Penalty Kick game as well. Table 11 reports the value of the Spearman rank correlation coefficient between the first (i.e., 75 of 150 plays) and second half frequencies of *Right*.[18]

	Professionals			Students		
	R	t	p-value	R	t	p-value
Row	-0.4153	-1.94	0.06821	-0.0982	-0.41	0.68053
Column	-0.4354	-2.05	0.05522	-0.2155	-0.94	0.36153

Table 11: Spearman Rank Correlation Coefficients

Since there are only 20 professionals in each role, this test has lower power for the Penalty Kick game than for the O'Neill game where (since the equilibrium mixture was the same for both roles) we could pool the data for all 40 row and column players. Nonetheless, for professionals the null of no correlation is rejected at the 10% level, and just fails to be rejected at the 5% level, while students exhibit no correlation between their first and second half choice frequencies for either role.

Table 12 provides weak evidence that professionals in the column role tend to switch from underplaying *Right* to overplaying it (and vice-versa), with 14 of the 20 column players switching.[19] Since professionals in the row role tend to overplay *Right* in both halves, the null that the frequencies are uniform is rejected. The same null is rejected for students in the row role; they tend to underplay *Right* in both halves.

[18] In the Penalty Kick game each player has only two actions (L and R) and hence the correlation coefficient is the same whether we consider the *Left* frequencies or the *Right* frequencies.

[19] Under the null, the probability of 14 or more switch is 0.0577, hence one just fails to reject at the 10% level the null that professionals in the column role are equally likely to switch as not.

	1^{st} Half/ 2^{nd} Half	Under/ Over	Under/ Under	Over/ Over	Over/ Under	Total	Q	p-value
Pro.	Row	**1**	1	12	**6**	20	16.40	0.0009
	Column	**6**	3	3	**8**	20	3.60	0.3080
Student	Row	**6**	12	2	**0**	20	16.80	0.0008
	Column	**5**	9	1	**5**	20	6.40	0.0937

Table 12: Counts by Quadrant

Table 13 shows the results of applying the KS test to the distributions of p-values obtained when the randomized binomial test is applied to the Left-Right choices (over all 150 rounds) in the Penalty Kick game. For professionals, minimax play is rejected for both the row and the column roles. For the row role, the null is rejected as there are too many small p-values (14 of the p-values are 0.39 or smaller). For the column role there are too many large p-values (17 of the 20 p-values are 0.4872 or higher).

	Professionals			Students	
	KS	p-value		KS	p-value
Row	1.403	0.039058	Row	2.732	0.000001
Column	1.508	0.021187	Column	2.509	0.000007

Table 13: Overall – KS tests of conformity to $U[0, 1]$

Minimax play is resoundingly rejected by the KS test for students, with too many small p-values for both the row and the column role.[20]

[20] In contrast to the O'Neill game, in the Penalty kick game the behavior of professionals conforms more closely to equilibrium at the individual level than does the behavior of students even when each half of the data is considered in isolation. In the first half, the minimax binomial model is rejected at the 5% level for 3 professionals and 9 students. In the second half, minimax play is rejected for 3 professionals and 6 students. As in the O'Neill game, minimax play at the individual level is rejected more frequently for professionals in each half of the data than it is in the data overall, which provides further evidence that professionals follow different mixtures in each half.

Conclusion

We have shown that the behavior of professionals departs from the minimax hypothesis in a number of respects in both the O'Neill game and the Penalty Kick game – there is negative correlation between the first and second half choice frequencies of each action. In addition, in the O'Neill game there is a striking tendency for professionals to switch from underplaying a card relative to its equilibrium frequency to overplaying it (and vice-versa). Negative correlation and switching cause the empirical choice frequencies to be too close to the theoretical frequencies for each of the four cards in the O'Neill game and for the column player in the Penalty kick game. Applying the KS test to the distribution of p-values (for the overall data) for professionals, we strongly reject the minimax binomial model for each card individually and the minimax multinomial model for all cards jointly in the O'Neill game. The KS test also rejects minimax play for both the row and the column player in the Penalty Kick game.

Students, however, exhibit no correlation between their first-half and second-half choice frequencies in either game (except for the *Joker* card in the O'Neill game), nor do they exhibit any tendency to switch between under and overplaying a card. The KS test does not reject minimax play in the O'Neill game, although it is does in the Penalty Kick game. Hence, with respect to (i) correlation between halves, (ii) the likelihood of "switching" between halves, and (iii) the empirical distribution of p-values, the behavior of students conforms more closely to theory than does the behavior of professionals, especially for the O'Neill game.

Considering the overall data (200 rounds in O'Neill and 150 rounds in the Penalty Kick game) the empirical choices frequencies are closer to the theoretical predictions for professionals than students, but the difference seems to be a result of the fact that professionals do *not* follow the same *iid* mixture as the theory says they should. If in the PH-V experiment the O'Neill game had ended after 100 rounds instead of 200, their experiment would have found no significant difference between professionals and students in terms of the conformity of the their empirical mixtures to the minimax mixture. Nor have professionals learned to play minimax in the last 100 plays while

students have not – as we have shown, the behavior of students and professionals conforms equally well to minimax when the second half of the O'Neill game data is considered alone.

Why do professional follow non-stationary mixtures, switching from overplaying a card to underplaying it (or vice-versa)? Perhaps professionals initially play a non-minimax mixture with the goal of inducing their rival to play a non-minimax mixture and then subsequently exploiting their rival's deviation. For example, a row player might overplay the *Joker* card in an attempt to get his rival to overplay *non-Joker* cards and then exploit his rival by switching to overplaying *non-Joker* cards.[21] Such behavior is, of course, inconsistent with the minimax hypothesis; moreover, it is unclear why, except by chance, it would lead to choice frequencies over all 200 rounds that are close to the minimax frequencies.

Alternatively, perhaps there was some flaw in the conduct of the experiment. To a professional who has overplayed *Joker* in the first half of the experiment, the experimenter might have inadvertently provided a cue to play *Joker* less frequently in the second half. In a mixed-strategy equilibrium, a player is indifferent between alternative actions and hence even a small cue or intimation might have a significant influence on choices. (The cue need not even be consciously noticed by the subject or the experimenter.) This might explain the negative correlation and tendency to switch that we have documented for professionals.

Since there is no obvious theoretical justification why professionals (but not students) would follow non-stationary mixtures, efforts to replicate the PH-V results seem especially important. Our results suggest that it may be useful to focus on the whether professionals follow non-stationary mixtures in the analysis of replication studies. Although not an exact replication, Levitt, List, and Reiley (2007) find no evidence that the behavior of American Major League Soccer players conforms more closely to minimax than does the behavior of students in the O'Neill game. Van Essen and Wooders (2007) find differences in the behavior of experienced and novice poker players in

[21] I'm grateful to Mark Walker for suggesting this possibility.

a mixed-strategy poker game, but even experienced players generally do not play in conformity with minimax.

References

[1] Brown, D. and R. Rosenthal (1990): "Testing the Minimax Hypothesis: A Reexamination of O'Neill's Experiment," *Econometrica* **58**, pp. 1065-1081.

[2] Camerer, C. (2003): *Behavioral Game Theory, Experiments in Strategic Interaction*, Princeton University Press, Princeton.

[3] Chiappori, P., S. Levitt, and T. Groseclose (2002): "Testing Mixed Strategy Equilibria When Players are Heterogeneous: The Case of Penalty Kicks in Soccer," *American Economic Review* **92**, pp. 1138-1151.

[4] Hsu, S., Huang, C. and C. Tang (2007): "Minimax Play at Wimbledon: Comment," *American Economic Review* **97**, pp. 517-523.

[5] Gibbons, J. and S. Chakraborti (2003): *Nonparametric Statistical Inference*, New York: Marcel Dekker.

[6] Levitt, S., List, J., and D. Reiley (2007): "What Happens in the Field Stays in the Field: Professionals Do Not Play Minimax in Laboratory Experiments," University of Arizona working paper 07-11.

[7] Mood, A., Graybill, F., and D. Boes (1974). *Introduction to the Theory of Statistics*, New York: McGraw Hill.

[8] Mookherjee, D. and B. Sopher (1994): "Learning behavior in an experimental matching pennies game," *Games and Economic Behavior* **7**, 62-91.

[9] Ochs, J. (1994): "Games with a Unique Mixed Strategy Equilibria: An Experimental Study," *Games and Economic Behavior* **10**, 202-217.

[10] O'Neill, B. (1987), "Nonmetric Test of the Minimax Theory of Two-Person Zero-Sum Games," *Proceedings of the National Academy of Sciences* **84**, 2106-2109.

[11] O'Neill, B. (1991). "Comments on Brown and Rosenthal's Reexamination," *Econometrica* **59**, 503-507.

[12] Palacios-Huerta, I. (2003): "Professionals Play Minimax," *Review of Economic Studies* **70**, pp. 395-415.

[13] Palacios-Huerta, I. and O. Volij (2008): "Experientia Docent: Professionals Play Minimax in Laboratory Experiments," *Econometrica* **76**, pp. 71-115.

[14] Rapoport, A. and R. Boebel (1992): "Mixed Strategies in Strictly Competitive Games: A Further Test of the Minimax Hypothesis," *Games and Economic Behavior* **4**, 261-283.

[15] Rosenthal, R., J. Shachat, and M. Walker (2003): "Hide and seek in Arizona," *International Journal of Game Theory* **32**, pp. 273-293.

[16] Shachat, J. (2002): "Mixed Strategy Play and the Minimax Hypothesis," *Journal of Economic Theory* **104**, pp. 189-226.

[17] Walker, M. and J. Wooders (2001): "Minimax Play at Wimbledon," *American Economic Review* **91**, pp. 1521-1538.

[18] Van Essen, M. and J. Wooders (2007): "Blind Stealing: Experience and Expertise in a Mixed Strategy Poker Experiment," working paper.

[19] Wooders, J. and J. Shachat (2001): "On The Irrelevance of Risk Attitudes in Repeated Two-Outcome Games," *Games and Economic Behavior* **34**, pp. 342-363.

Table 1: Professionals Playing O'Neill

		First Half				Second Half				Overall			
Pair	Player	1	2	3	J	1	2	3	J	1	2	3	J
1	C	23	23	24	30 **	16	14	18	52 **				
	R	19	28 *	24	29 **	19	17	34 **	30 **			**	**
2	C	20	22	23	35	20	19	27 *	34			*	*
	R	23	22	22	33	18	21	27	34				*
3	C	25	21	23	31 *	14	14	18	54 **				
	R	25	25	14	36	17	14	26	43				
4	C	18	20	21	41	11 **	17	24	48	**			
	R	19	19	18	44	24	22	18	36				
5	C	22	21	16	41	18	18	26	38				
	R	20	17	28 **	35	16	22	13 *	49 *				
6	C	27 *	18	17	38	14	19	24	43				
	R	21	20	16	43	21	21	21	37				
7	C	23	21	19	37	18	17	22	43				
	R	20	24	13 *	43	23	19	13 *	45			**	
8	C	26	15	24	35	19	15	17	49 *		*		
	R	22	23	19	36	17	20	20	43				
9	C	20	16	20	44	21	20	21	38				
	R	22	21	21	36	15	18	22	45				
10	C	21	25	16	38	18	14	27 *	41				
	R	14	17	7 **	62 **	21	19	27	33				**
11	C	27 *	22	23	28 **	23	18	18	41	*			
	R	21	21	12 **	46	20	17	22	41				
12	C	17	23	20	40	22	17	21	40				
	R	16	18	23	43	24	22	16	38				
13	C	21	23	17	39	18	20	21	41				
	R	17	21	18	44	26	16	21	37				
14	C	20	31 **	22	27 **	21	27 *	17	35		**		**
	R	17	12 **	18	53 **	20	25	23	32				
15	C	21	23	24	32 *	21	14	16	49 *				
	R	19	19	13 *	49 *	24	21	21	34				
16	C	23	20	16	41	16	13 *	19	52 **				*
	R	17	19	14	50 **	24	20	25	31 *				
17	C	22	27	18	33	23	16	23	38				
	R	16	28 *	21	35	25	18	17	40				
18	C	20	29 **	19	32 *	21	20	23	36				*
	R	25	16	17	42	17	23	19	41				
19	C	21	21	18	40	13 *	20	17	50 **				
	R	19	20	18	43	22	24	29 **	25 **				*
20	C	22	21	21	36	15	20	15	50 **				
	R	21	19	21	39	18	23	19	40				
	C	439 **	442 **	401	718 **	362 **	352 **	414	872 **				
	R	393	409	357 **	841 *	411	402	433 *	754 **				
	Overall	832	851 **	758 *	1559	773	754 *	847 *	1626				

Notes: ** and * denote rejection of minimax binomial model for a given card at the 5% and 10% level, respectively.

Table 2: Students Playing O'Neill

		First Half				Second Half				Overall			
Pair	Player	1	2	3	J	1	2	3	J	1	2	3	J
1	C	16	19	23	42	12 **	22	29 **	37	**		**	
	R	19	28 **	22	31 *	26	26	17	31 *		**		**
2	C	15	19	17	49 *	22	22	25	31 *				
	R	19	16	14 *	51 **	22	20	18	40				
3	C	20	21	20	39	26	17	25	32 *				
	R	16	24	21	39	24	17	22	37				
4	C	17	18	20	45	18	18	25	39				
	R	13 *	16	15	56 **	16	27 *	16	41	*			**
5	C	17	20	17	46	22	15	22	41				
	R	11 **	19	22	48	16	19	25	40	**			
6	C	25	24	21	30 **	18	22	25	35				**
	R	16	19	27 *	38	21	24	20	35				
7	C	20	12 **	19	49 *	20	18	14	48		*		**
	R	21	20	22	37	25	17	21	37				
8	C	20	20	15	45	17	25	22	36				
	R	22	21	17	40	17	24	16	43				
9	C	15	22	19	44	25	21	20	34				
	R	14	19	24	43	16	24	16	44	*			
10	C	27 *	15	25	33	23	22	17	38	*			
	R	33 **	27 *	20	20 **	23	25	20	32 *	**	**		**
11	C	15	28 *	21	36	30 **	24	20	26 **		**		**
	R	19	14	19	48	20	21	17	42				
12	C	20	19	13 *	48	23	25	22	30 **				
	R	25	21	17	37	31 **	21	19	29 **	**			**
13	C	16	18	17	49 *	24	22	25	29 **				
	R	17	26	18	39	18	13 *	21	48 *				
14	C	19	19	16	46	20	20	25	35				
	R	18	20	23	39	16	26	29 **	29 **			**	*
15	C	17	21	17	45	22	20	23	35				
	R	14	19	17	50 **	14	23	23	40	**			
16	C	23	21	17	39	22	22	15	41				
	R	21	22	22	35	28 *	17	16	39				
17	C	19	21	18	42	23	20	22	35				
	R	22	14	20	44	17	18	20	45				
18	C	25	21	17	37	18	20	19	43				
	R	26	21	20	33	27 *	21	17	35	**			*
19	C	28 *	25	19	28 **	23	23	20	34	*			**
	R	25	19	24	32 *	35 **	18	14	33	**			**
20	C	24	21	17	38	19	18	24	39				
	R	24	16	14	46	15	25	19	41				
	C	398	404	368 *	830	427	416	439 **	718 **				*
	R	395	401	398	806	427	426	386	761 *				
	Overall	793	805	766	1636	854 **	842 *	825	1479 **				*

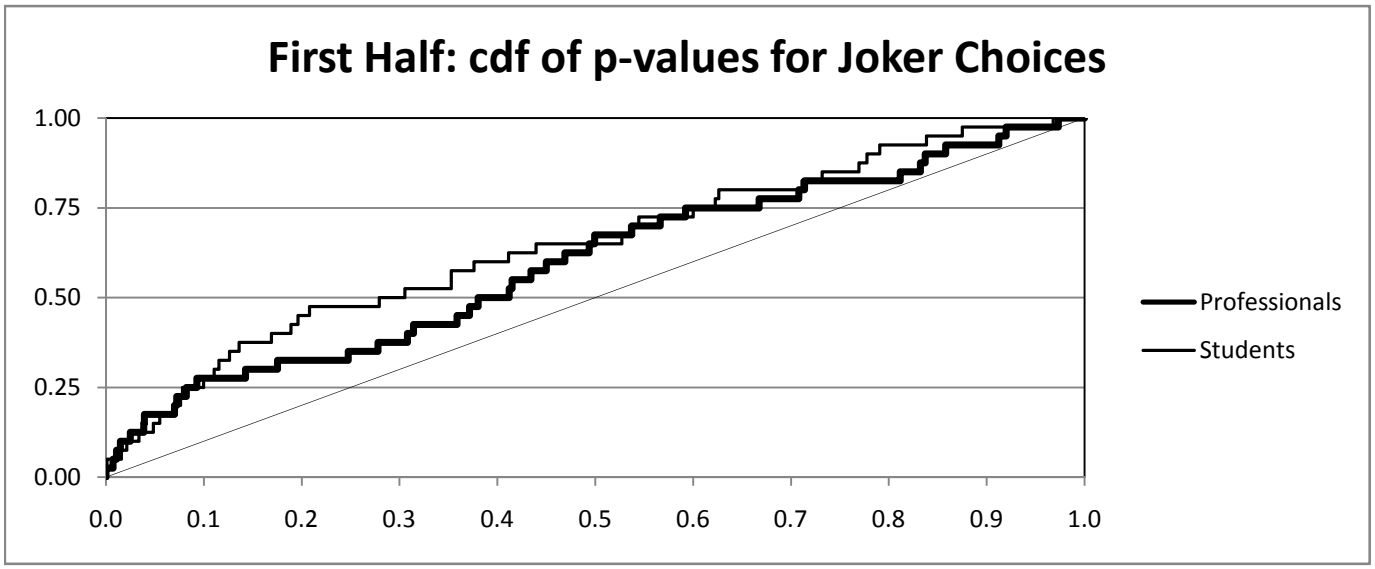
Notes: ** and * denote rejection of minimax binomial model for a given card at the 5% and 10% level, respectively.

Table 3: Pearson Goodness of Fit, All Cards

		Professionals			Students		
Pair	Player	First 100	Last 100	Overall	First 100	Last 100	Overall
1	C	0.241	0.094 *	0.940	0.706	0.053 *	0.065 *
	R	0.070 *	0.005 **	0.002 **	0.140	0.108	0.022 **
2	C	0.735	0.334	0.257	0.287	0.299	0.950
	R	0.557	0.308	0.223	0.129	0.940	0.316
3	C	0.287	0.034 **	0.804	0.995	0.165	0.451
	R	0.195	0.233	0.987	0.643	0.643	0.930
4	C	0.965	0.075 *	0.180	0.735	0.643	0.604
	R	0.873	0.659	0.885	0.012 **	0.253	0.031 **
5	C	0.783	0.513	0.987	0.615	0.643	0.732
	R	0.233	0.140	0.885	0.117	0.552	0.096 *
6	C	0.362	0.411	0.962	0.204	0.517	0.184
	R	0.783	0.945	0.950	0.334	0.688	0.515
7	C	0.855	0.783	0.985	0.153	0.308	0.062 *
	R	0.324	0.311	0.105	0.924	0.578	0.615
8	C	0.215	0.287	0.341	0.599	0.513	0.780
	R	0.777	0.879	0.962	0.873	0.517	0.575
9	C	0.753	0.978	0.919	0.593	0.532	0.960
	R	0.873	0.517	0.922	0.411	0.423	0.342
10	C	0.532	0.215	0.962	0.103	0.753	0.280
	R	0.000 **	0.287	0.192	0.000 **	0.348	0.000 **
11	C	0.082 *	0.831	0.257	0.179	0.013 **	0.040 **
	R	0.241	0.879	0.651	0.321	0.896	0.513
12	C	0.825	0.873	0.997	0.251	0.221	0.729
	R	0.643	0.593	0.998	0.578	0.027 **	0.025 **
13	C	0.807	0.965	0.950	0.324	0.153	0.985
	R	0.777	0.411	0.922	0.480	0.231	0.732
14	C	0.015 **	0.311	0.007 **	0.615	0.599	0.993
	R	0.044 **	0.348	0.852	0.879	0.022 **	0.066 *
15	C	0.407	0.197	0.953	0.665	0.735	0.997
	R	0.206	0.615	0.744	0.187	0.440	0.175
16	C	0.735	0.075 *	0.263	0.807	0.643	0.484
	R	0.187	0.253	0.993	0.783	0.215	0.457
17	C	0.253	0.615	0.596	0.940	0.735	0.971
	R	0.197	0.593	0.720	0.494	0.735	0.451
18	C	0.127	0.825	0.267	0.578	0.924	0.885
	R	0.457	0.807	0.888	0.380	0.311	0.096 *
19	C	0.960	0.145	0.423	0.044 **	0.615	0.034 **
	R	0.924	0.014 **	0.327	0.296	0.002 **	0.004 **
20	C	0.873	0.172	0.777	0.706	0.783	0.943
	R	0.982	0.873	0.987	0.231	0.462	0.596
	C	0.606	0.173	0.913	0.59	0.43	0.72
	R	0.041 **	0.242	0.962	0.02 **	0.05 *	0.00 **
	Overall	0.140	0.131	0.988	0.08 *	0.10	0.006 **

Notes: ** and * denote rejection of minimax multinomial model for a given card at the 5% and 10% level, respectively.

Figure 1: KS test on First Half for Joker Choices and All Card Choices



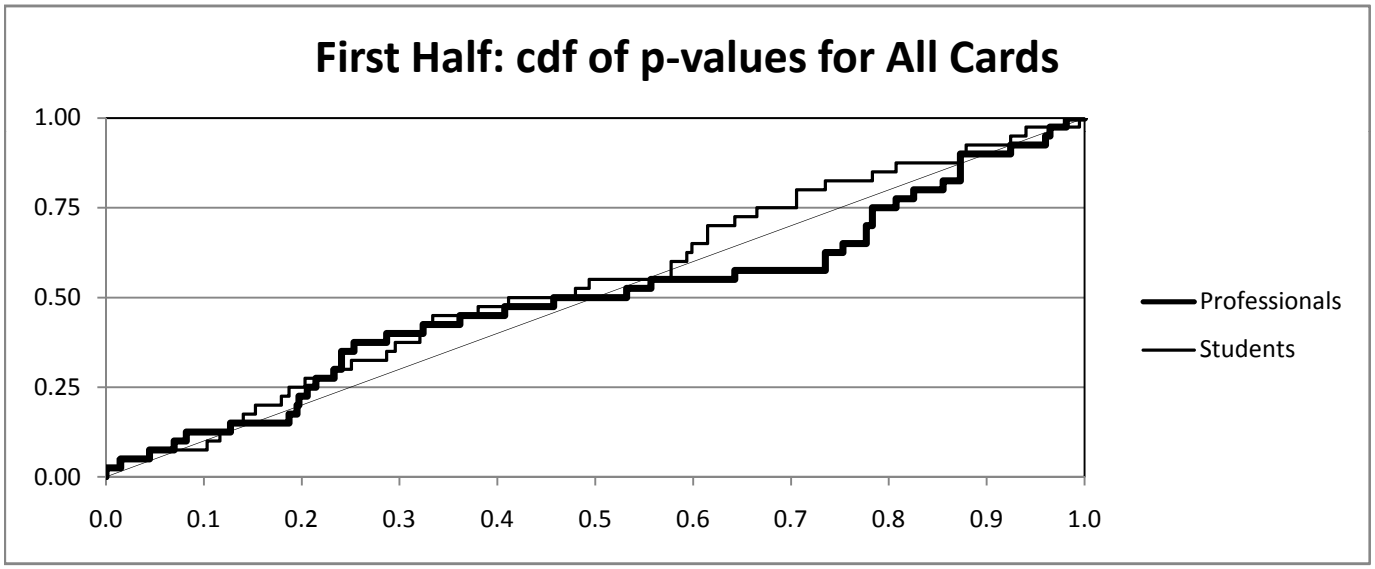
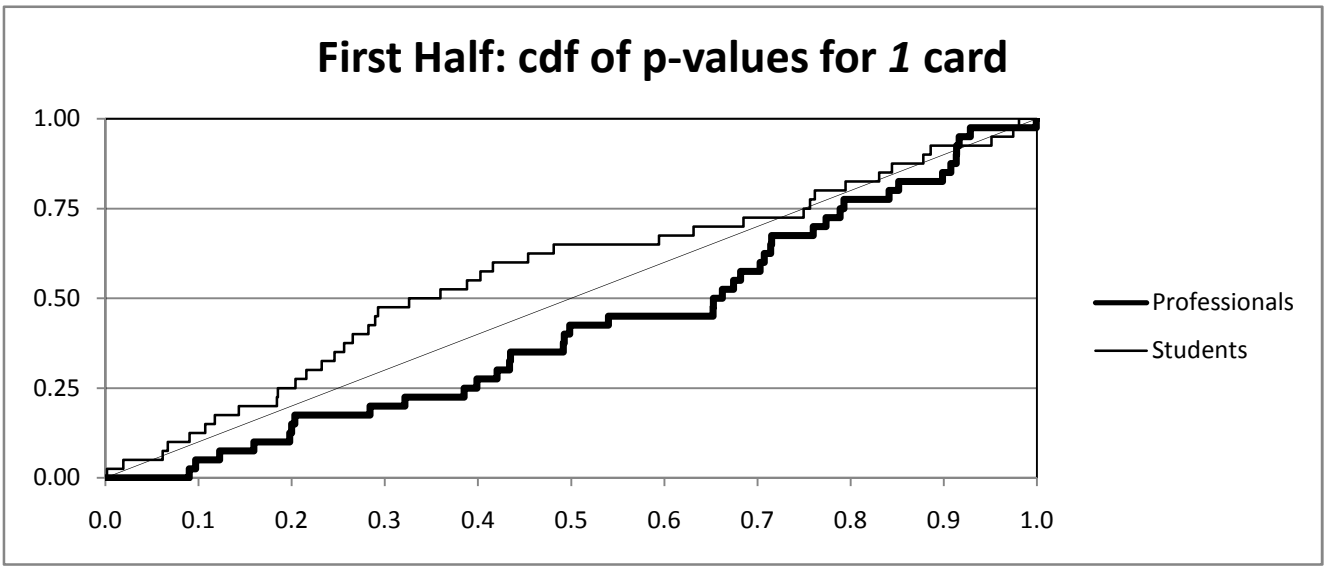
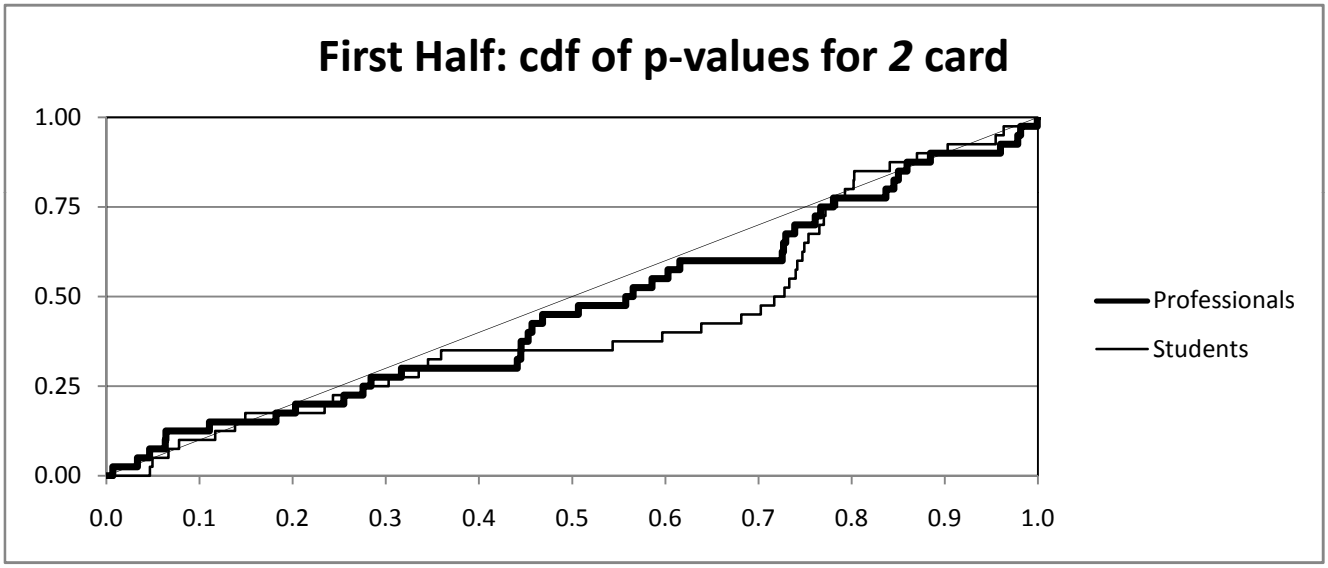


Figure 2: KS test on First Half for non-Joker Cards





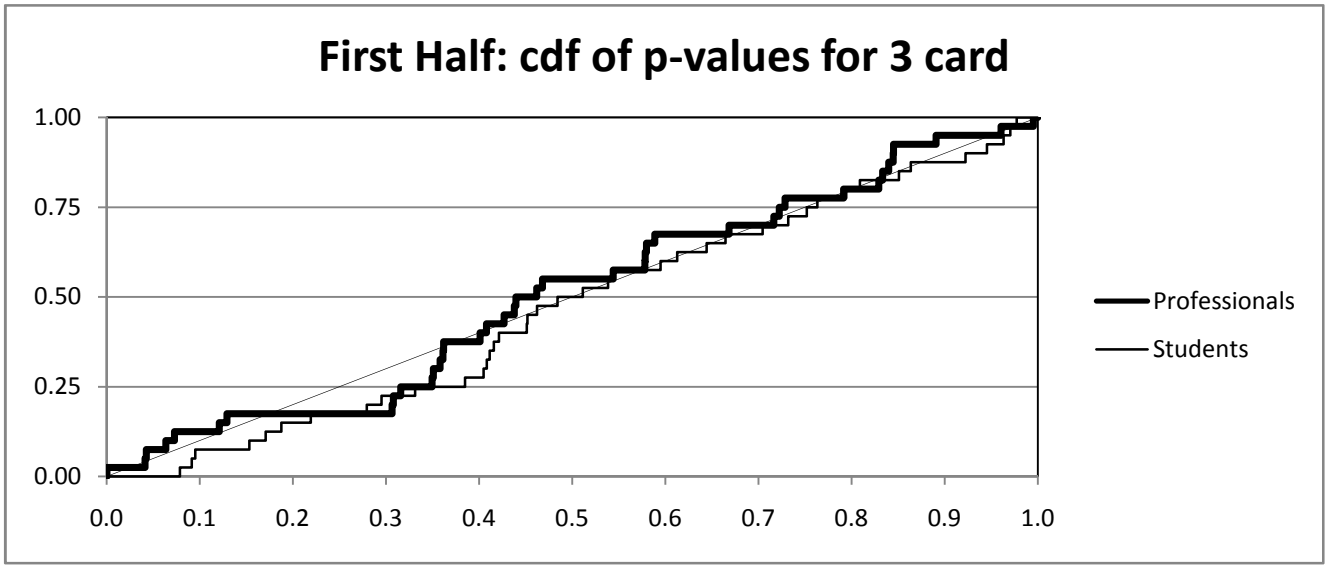
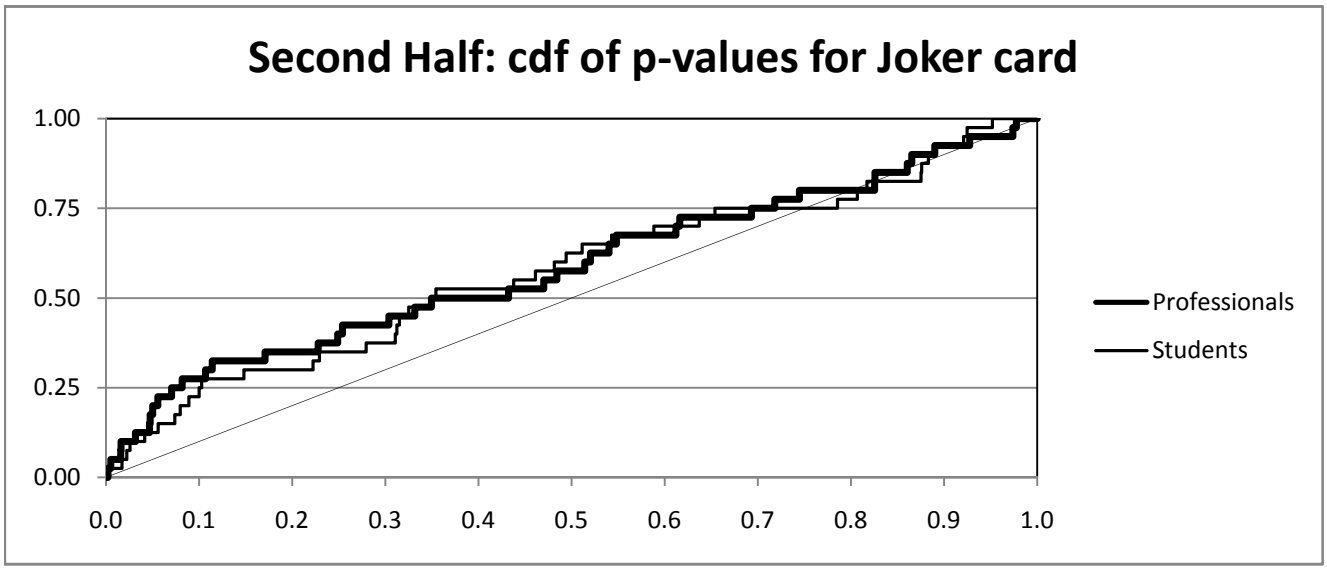


Figure 3: KS test on Second Half for Joker Choices and All Card Choices



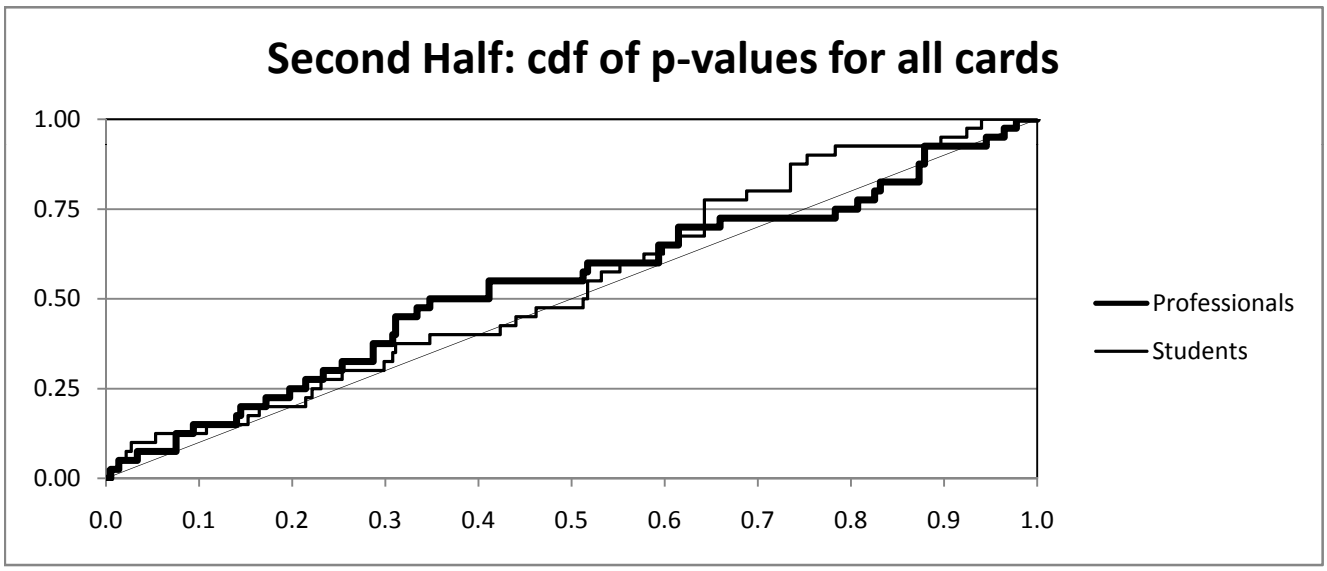
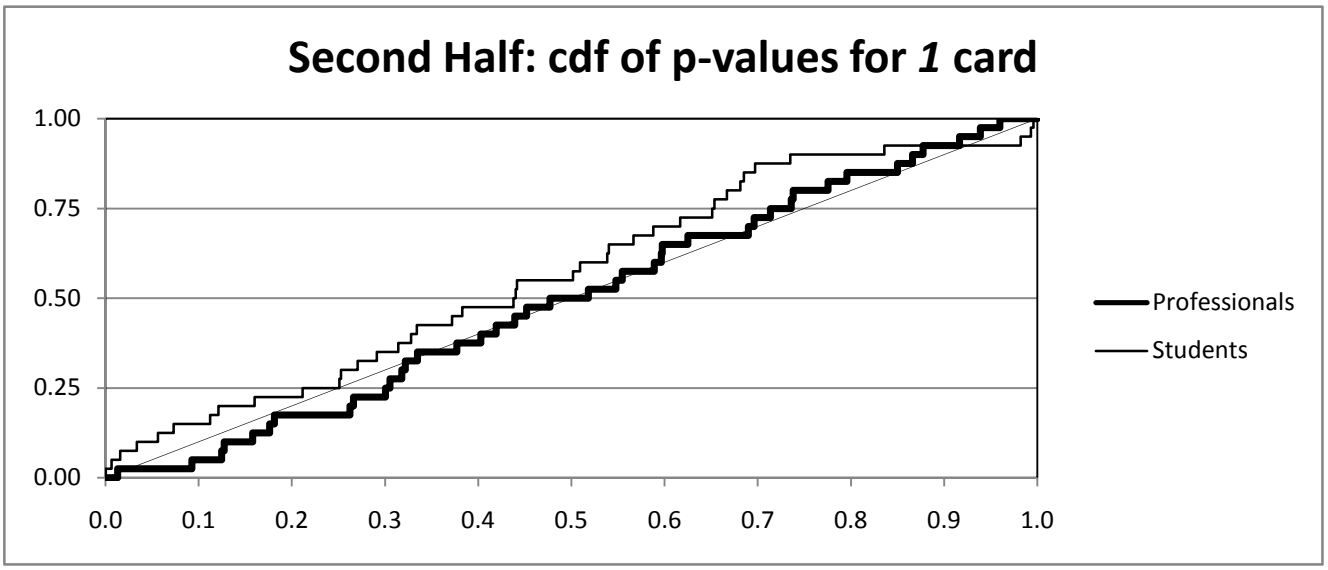
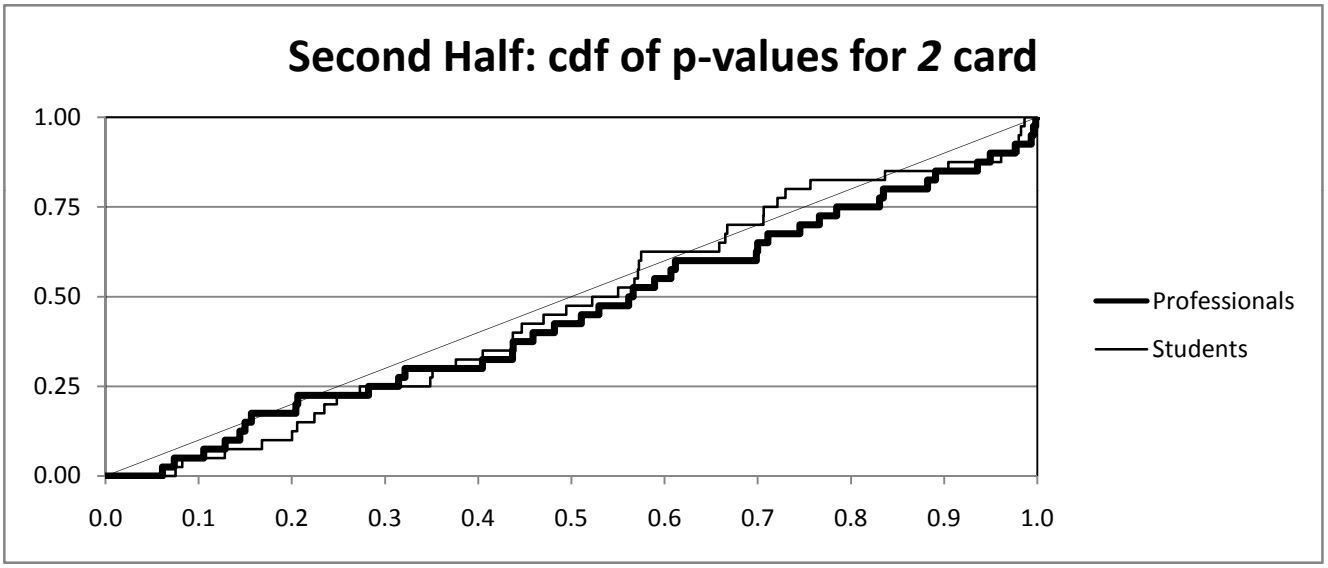


Figure 4: KS test on Second Half for non-Joker Cards





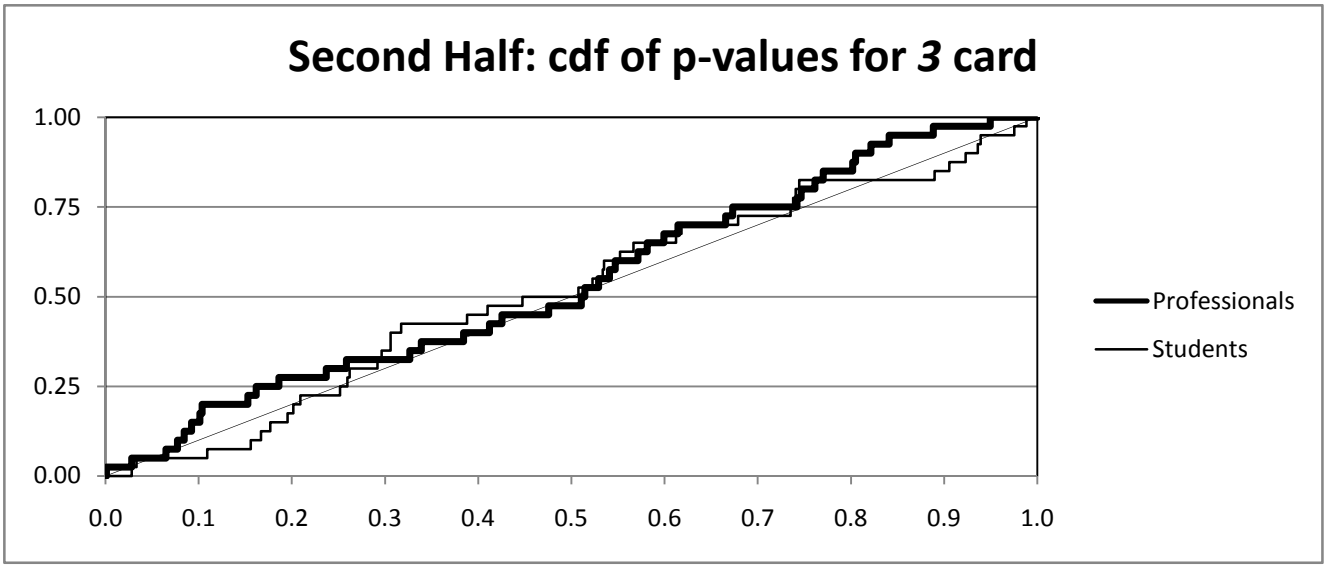
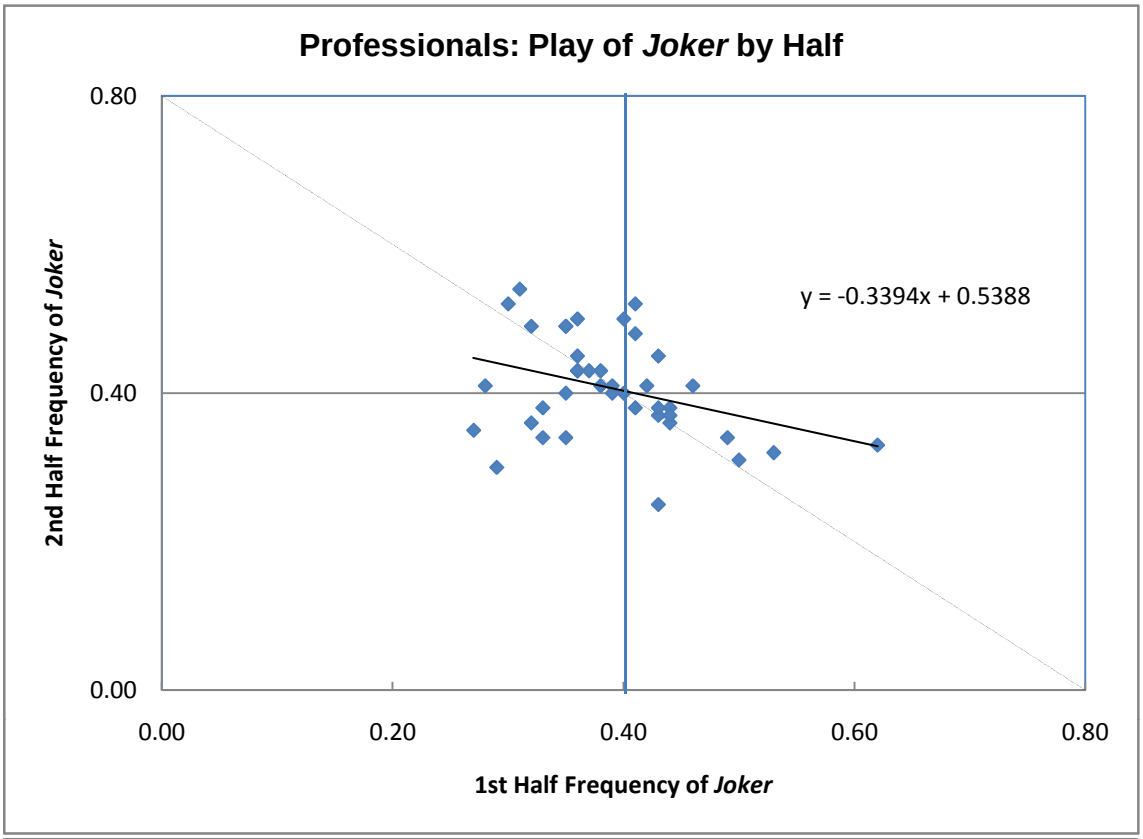


Figure 5: Joker Frequencies By Half



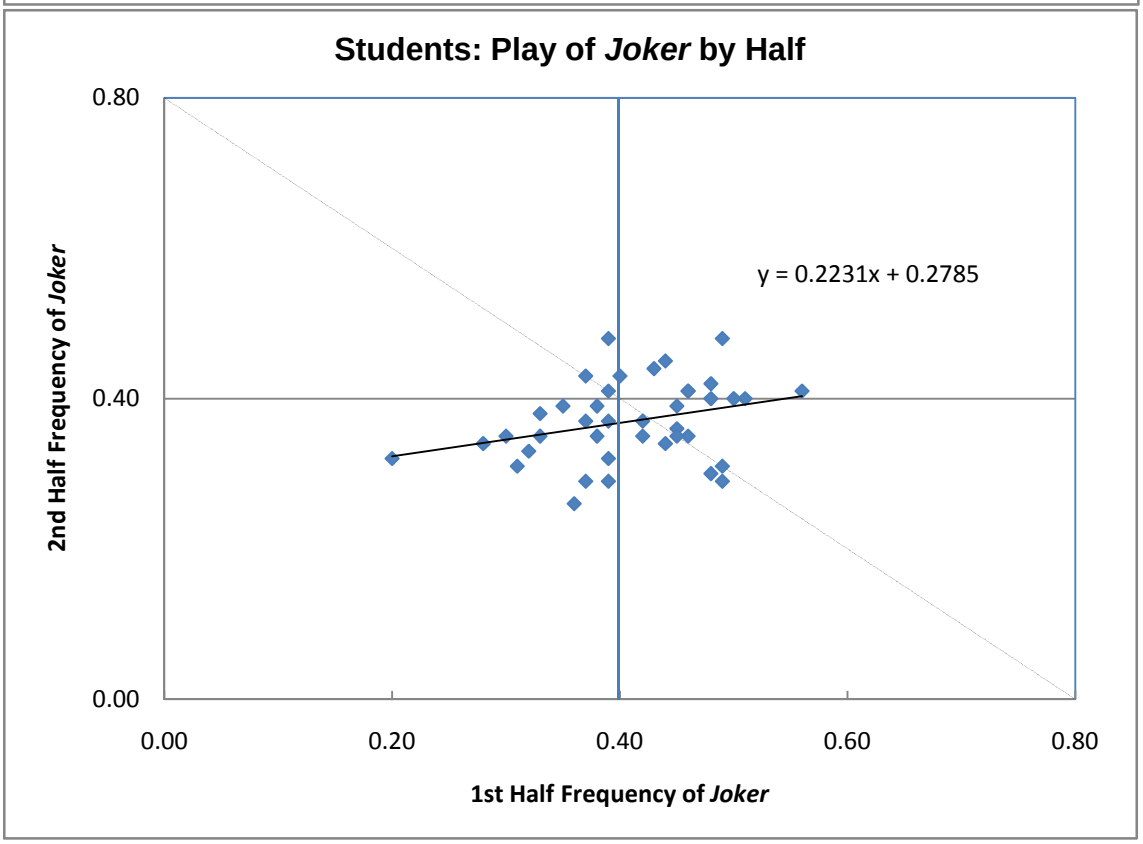
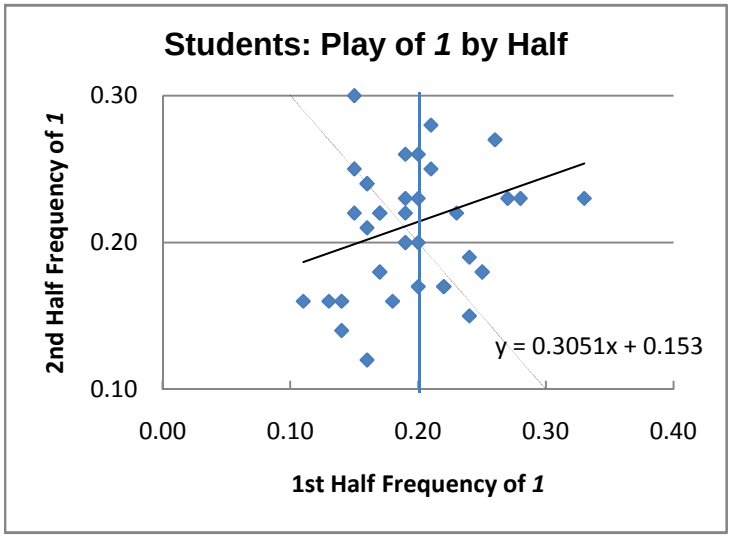
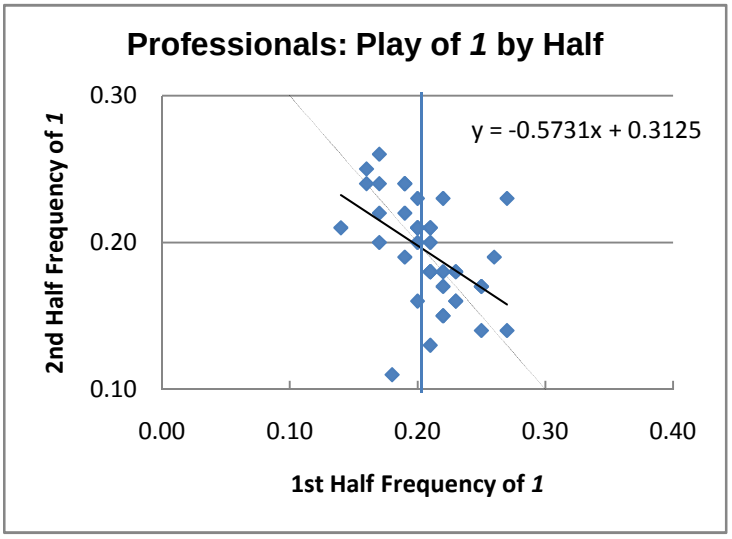
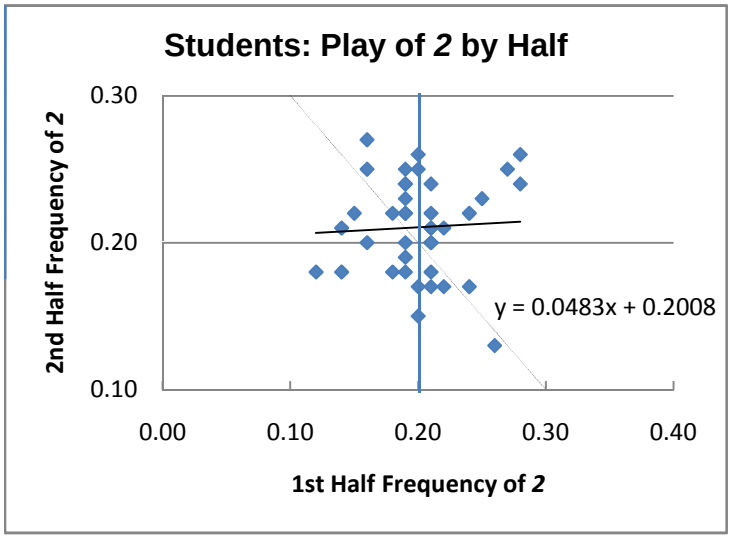
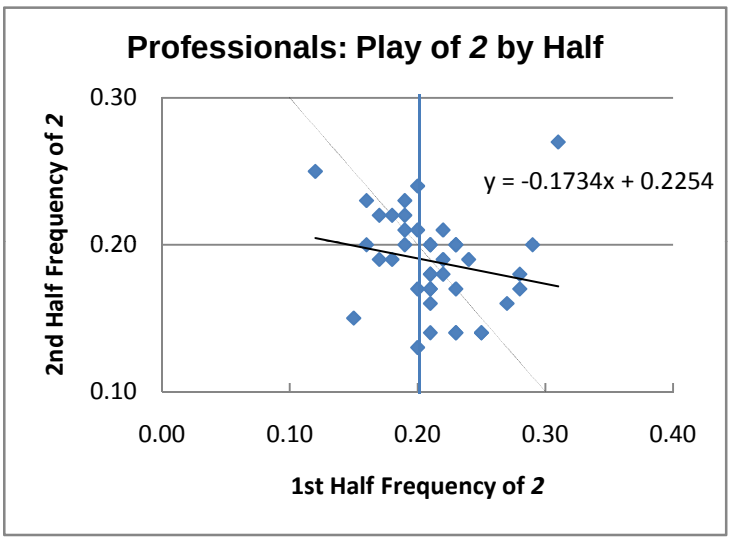


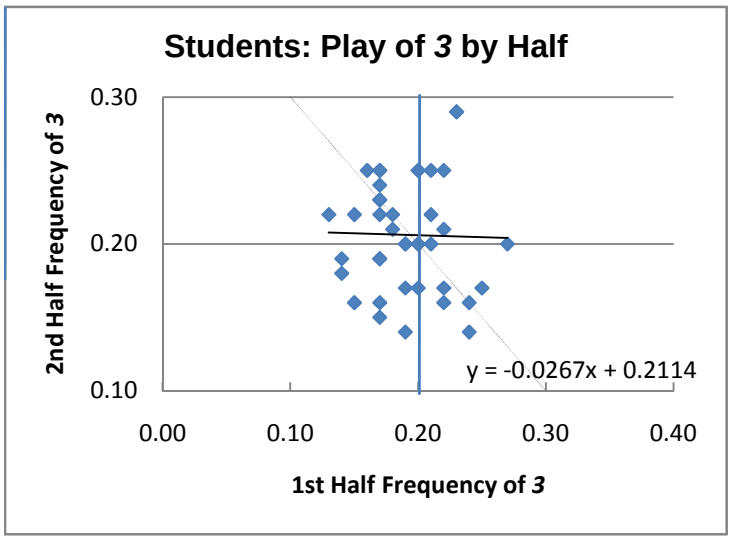
Figure 6: *non-Joker* Frequencies By Half











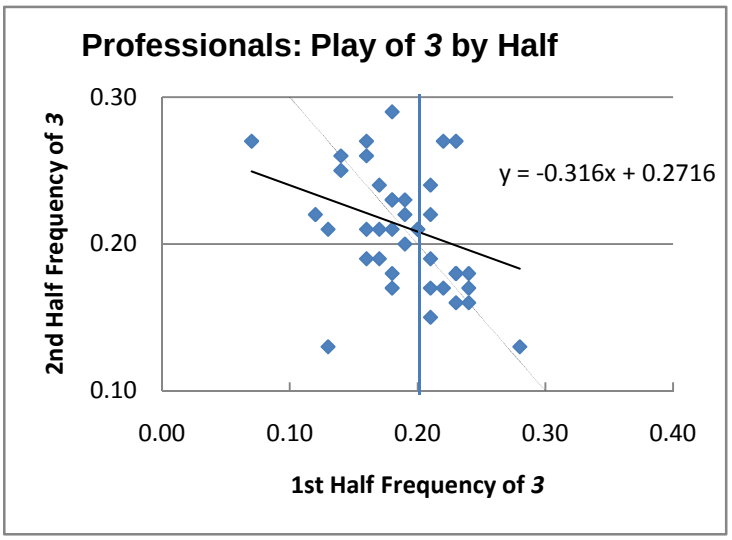


Figure 7: KS test on Overall Play for Joker Cards and All Cards Jointly

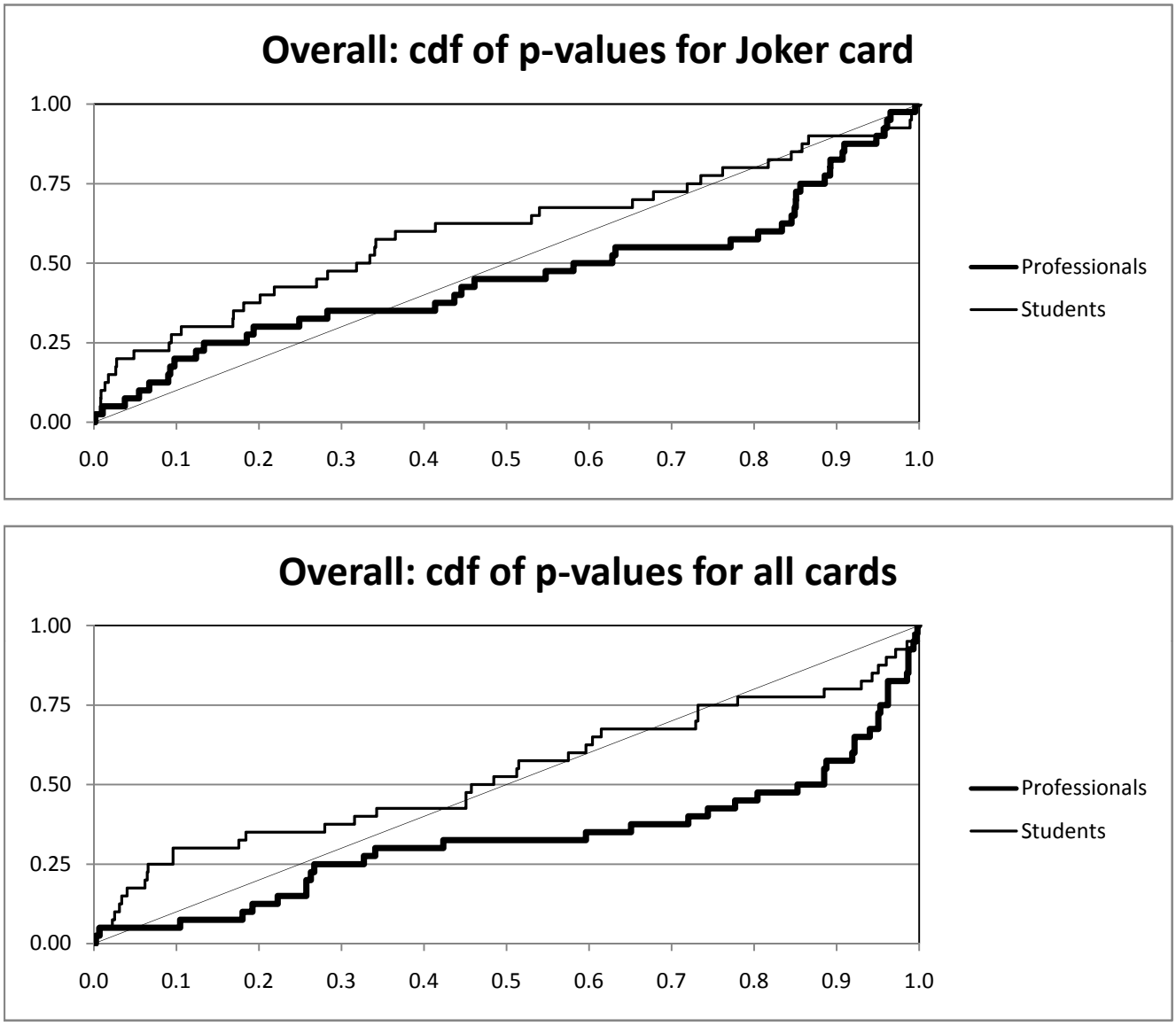


Figure 8: KS test on Overall Play for non-Joker Cards