

4 Teste piloto com agentes humanos

A consecução do objetivo do nosso trabalho depende da premissa fundamental de que agentes humanos não-especialistas são capazes, dentro de certos limites, de diferenciar perfis de fraudadores dos perfis de vendedores normais. Para obter indícios acerca da validade dessa premissa, executamos um teste piloto com um grupo de pessoas. Disponibilizamos um conjunto de perfis de vendedores e lhes solicitamos que respondessem, para cada perfil, se o consideravam pertencente a um fraudador ou a um vendedor normal. Em seguida, comparamos essas respostas com os desfechos reais *futuros*, dados pelo aparecimento ou não de indícios de fraude de acordo com os critérios apresentados na seção 3.3. Com isso, pudemos avaliar o quanto os agentes humanos previram corretamente o comportamento futuro dos vendedores analisados.

Nas próximas seções vamos detalhar as métricas utilizadas para avaliar o desempenho dos agentes humanos, a questão dos incentivos, o projeto do teste piloto, os resultados obtidos e a análise desses resultados.

4.1. Medindo o desempenho dos agentes humanos

Para que o jogo funcione, os agentes humanos devem classificar os vendedores de forma a agregar algum tipo de informação. Vamos agora operacionalizar esse critério. Dado um agente humano, sejam:

- VE o conjunto de vendedores a serem analisados.
- VE_F a partição de VE contendo fraudadores.
- VE_N a partição de VE contendo vendedores normais.
- $VE_F \cup VE_N = VE$ e $VE_F \cap VE_N = \emptyset$
- VE_{CN} o conjunto de vendedores que o agente humano classificou como vendedores normais.
- VE_{CF} o conjunto de vendedores que o agente humano classificou como fraudadores.

- $VE_{CN} \cup VE_{CF} = VE$ e $VE_{CN} \cap VE_{CF} = \emptyset$.
- $VE_{FCF} = VE_{CF} \cap VE_F$, isto é, o conjunto de fraudadores classificados como fraudadores (verdadeiros positivos).
- $VE_{NCF} = VE_{CF} \cap VE_N$, isto é, o conjunto de vendedores normais classificados como fraudadores (falsos positivos).

Com esses conjuntos definiremos duas métricas para a atuação do agente em questão: FCF (fração dos fraudadores classificados como fraudadores) e NCF (fração dos vendedores normais classificados como fraudadores).

$$\begin{aligned} FCF &= \frac{|VE_{FCF}|}{|VE_F|} \\ NCF &= \frac{|VE_{NCF}|}{|VE_N|} \end{aligned} \tag{4.1}$$

O que interessa para o jogo é que $FCF > NCF$, isto é, que o agente humano classifique como fraudadores proporcionalmente *mais* fraudadores que vendedores normais. Com isso, ele estará produzindo um conjunto VE_{CF} que contém proporcionalmente mais fraudadores que o conjunto original VE , contribuindo para o jogo.

Essas medidas (FCF e NCF) serão utilizadas na seção 5.5 como estimadores das seguintes probabilidades associadas a cada agente humano:

- Probabilidade de classificar como fraudador um vendedor, dado que de fato ele é um fraudador
- Probabilidade de classificar um vendedor normal como fraudador.

A precisão dessas estimativas dependerá da quantidade de vendedores de cada tipo (normal ou fraudador) que forem avaliados.

4.2. A questão dos incentivos

A proposta final deste trabalho é a utilização do mecanismo de classificação de vendedores junto com um mecanismo de incentivo aos agentes humanos, de modo que esses agentes sintam-se motivados a desempenhar bem sua tarefa e a buscar subsídios para isso (analisando seus acertos e erros passados, por exemplo). O “custo” de executar a tarefa pode ser compensado com incentivos que premiem aqueles que obtêm melhores resultados. Não se trata apenas de compen-

sar os incentivos para responder errado ou aleatoriamente, como acontece nos sistemas de recomendação (Ekstrom, Garcia & Bjornsson, 2005; Jurca, 2007; Miller et al., 2005): é necessário compensar o custo de execução da tarefa.

Um teste piloto ideal teria que seguir essa lógica: atrelar algum tipo de incentivo ao desempenho da tarefa e permitir que os participantes tivessem as mesmas opções de atuação que um agente humano teria com o jogo: busca de informações na Internet sobre os vendedores analisados e sobre fraudes em geral, possibilidade de discutir com outras pessoas etc. No entanto, por limitações de ordem prática, optamos por fazer um teste piloto mais simples, sem mecanismos de incentivo. Para compensar a ausência de incentivos, partimos da premissa de que os maiores impactos da ausência de incentivo seriam os seguintes:

- Falta de um processo ativo de aprendizagem: os participantes utilizariam apenas os conhecimentos que porventura possuísem sobre fraudes, além da experiência adquirida ao longo da execução do teste piloto.
- Resistência a avaliar muitos vendedores: como se trata de uma tarefa que pode tomar algum tempo, o desempenho dos participantes poderia ser minado pelo cansaço ou falta de interesse na tarefa.

Como forma de contrabalançar esses dois elementos negativos, optamos por (i) incluir no estudo uma *fase de treinamento* e (ii) limitar a quantidade de vendedores a ser avaliada em função do tempo que as pessoas aceitariam gastar na tarefa.

Na fase de treinamento, cada participante recebeu informações sobre fraudes em mercados eletrônicos, exemplos de perfis de fraudadores, indícios de conduta fraudulenta, estratégias comumente utilizadas pelos fraudadores etc., de modo a “simular” o desempenho que ele alcançaria se tivesse estímulo para buscar essas informações.

Com relação ao tempo, estimamos que duas horas seriam suficientes para o desempenho da tarefa sem cansar excessivamente os participantes.

4.3. Projeto do teste piloto

O projeto do teste se deu em duas fases: fizemos uma primeira tentativa de projeto seguida de um piloto com uma pessoa apenas. A partir das observações

feitas nesse primeiro teste, fizemos as oportunas alterações. Nas próximas seções explicaremos como foi feita a seleção dos perfis de vendedores a serem analisados, o instrumento de coleta de dados utilizado e as instruções fornecidas na fase de treinamento. Daqui em diante vamos nos referir aos agentes humanos que participaram no teste piloto de *participantes*.

4.3.1. Perfis dos vendedores a serem analisados

Os participantes do teste piloto tinham como tarefa *prever* se o vendedor sendo analisado era um vendedor normal, que entregaria os produtos, ou se era um fraudador, que perpetraria fraude de não-entrega. Obtivemos a resposta real a essa questão com o passar do tempo, aplicando os mesmos critérios utilizados na pesquisa exploratória (cfr. seção 3.3):

- Se apareciam indícios de conduta fraudulenta (sinais de fraude junto com suspensão do vendedor), o mesmo era considerado *fraudador*.
- Se faltavam os indícios acima e apareciam sinais de normalidade (novas qualificações positivas), o vendedor era considerado *normal*.

Na Figura 7 apresentamos uma linha do tempo mostrando a relação entre os eventos acima e indicando quando é obtida a resposta acerca da conduta do vendedor. Também apontamos na figura um intervalo de tempo importante: a *antecedência* com que o participante do teste dá sua resposta *com relação ao momento da suspensão do vendedor*. Quanto maior a antecedência, melhor, já que interessa para o operador aplicar logo medidas que restrinjam o prejuízo que o fraudador pode causar. Assumimos que as qualificações negativas e a conseqüente suspensão do vendedor ocorrem sempre mais rápido que a chegada de qualificações positivas.

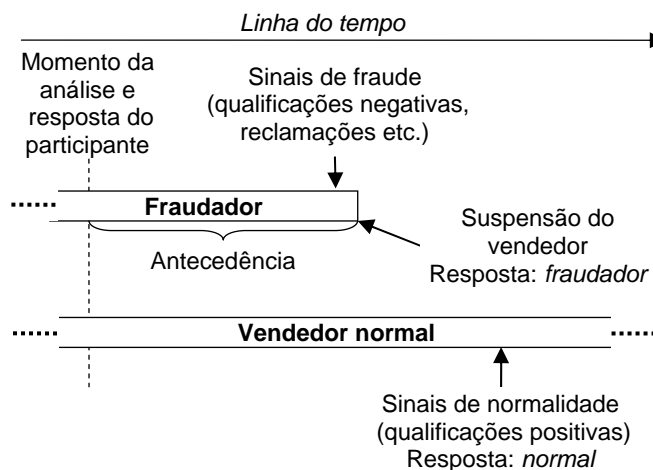


Figura 7: seqüência temporal dos eventos

A escolha dos perfis a serem analisados teve que atentar para algumas limitações de ordem prática:

- A quantidade de perfis deveria ser tal que poderia ser analisada dentro do tempo máximo previsto (2 horas).
- Deveria haver uma quantidade razoável de fraudadores, caso contrário não seria possível usar *FCF* como um estimador.
- Como os perfis variam ao longo do tempo (vendas são concretizadas, novas perguntas são feitas ao vendedor etc.), todos os participantes teriam que responder simultaneamente, de forma a garantir que eles estivessem diante das mesmas informações.

Para viabilizar o atendimento das duas primeiras restrições e relaxar a terceira, adotamos a solução de apresentar aos participantes “fotografias” (*snapshots*) dos perfis a serem analisados: de cada perfil fizemos uma cópia local de sua página no sítio do MercadoLivre. Além disso, também copiamos todas as páginas apontadas por ela e que seriam relevantes para a análise, e.g. páginas dos produtos à venda. Com isso, nós pudemos garantir que todos os participantes vissem as mesmas “fotografias” dos perfis dos vendedores, atendendo à terceira restrição.

Já para atender as duas primeiras restrições, a solução foi observar a evolução dos perfis “fotografados”, verificando quais apresentaram posteriormente sinais de fraude. Com isso, pudemos rotular cada uma das “fotografias” como pertencentes a um vendedor normal ou a um fraudador. Quando obtivemos uma quantidade aceitável de cópias de perfis de fraudadores, juntamos essas cópias com as de perfis de vendedores normais, produzindo então o conjunto a ser analisado pelos participantes.

A produção dessas cópias era muito custosa em termos de recursos de rede e precisávamos copiar todos os perfis para só depois saber quais seriam de fraudadores. Para reduzir a quantidade de perfis a que serem copiados, procuramos desenvolver um critério de seleção de perfis que maximizasse a quantidade de fraudadores. Estudamos os dados coletados na pesquisa exploratória sobre o MercadoLivre e desenvolvemos um critério tal que a maior parte dos perfis de fraudadores encontrados na pesquisa se encaixavam nesse critério *antes de serem inabilitados e receberem pontuações negativas*. Com isso, nós conseguimos aumentar a “concentração” de fraudadores no conjunto de perfis copiados.

As condições que utilizamos foram as seguintes:

- Vendedor estava habilitado, sem nenhum ponto negativo no mês e com total de pontos positivos menor que 400.
- As vendas nos últimos seis meses, excluídas as do último mês, eram menores que 100 (em quantidade).
- As ofertas do vendedor estavam há menos de 16 dias no sistema.
- Foram feitas mais de quatro vendas.
- O valor total à venda (quantidade de produtos multiplicados pelos seus valores) era maior que R\$ 3.000,00.
- Pelo menos um dos produtos anunciados pertencia a uma categoria na qual já tinham sido encontrados produtos de fraudadores na pesquisa exploratória.

Com isso, esperávamos delimitar uma população com maior probabilidade de albergar fraudadores. A cada dia, verificávamos que vendedores se enquadravam nessas condições e então fazíamos a cópia do seu perfil. No momento em que interrompemos a cópia de perfis, tínhamos 7 perfis de fraudadores e 59 perfis de vendedores normais. Ou seja, nossos critérios permitiram encontrar um conjunto de vendedores com pouco mais de 10% de fraudadores. É importante notar que essas cópias foram feitas vários dias antes (em média 5) dos fraudadores receberem qualificações negativas, simulando o que um usuário comum veria ao visitar o perfil do vendedor.

Estimamos que os participantes conseguiriam avaliar entre 20 e 30 vendedores em 2 horas. Montamos uma primeira lista de 30 vendedores, misturando os 7 perfis de fraudadores com 23 perfis de vendedores normais, escolhidos aleatoria-

mente dentre 59 vendedores normais identificados até o momento de elaboração da lista. A ordem dos vendedores na lista também foi aleatória.

Após a execução do teste com uma só pessoa, percebemos que a quantidade de vendedores ainda era grande e resolvemos reduzir para 20. Por outro lado, notamos que o participante estava inseguro nas primeiras respostas. Acrescentamos então 3 vendedores normais no começo do questionário para servirem de treinamento, sem informar os participantes disso. Então a lista final ficou com 23 vendedores, sendo que as três primeiras respostas foram desconsideradas nas estatísticas.

Os participantes acediam aos perfis dos vendedores através de uma página web contendo *links* para as cópias que fizemos. Para cada vendedor foi criado um link com seu apelido, apontando para o seu perfil de usuário na cópia. Dentro de cada perfil todos os *links* ou levavam para páginas locais ou então para uma página de erro, impedindo que o usuário “saísse” da cópia sem perceber e tomasse contato com os dados atuais dos vendedores em questão.

4.3.2. Instrumento de coleta de dados

O primeiro instrumento que fizemos foi uma tabela impressa com as seguintes colunas: o apelido do vendedor (aparecendo na mesma ordem que na página com os *links* para o espelho), a resposta do participante, o grau de certeza na resposta dada (de 1 a 5) e o tempo gasto para dar a resposta. Depois do teste com uma pessoa, resolvermos mudar o instrumento, passando a enviar a cada participante uma planilha do Google Documents¹⁵ (Figura 8) na qual seriam colocadas as respostas. Também reduzimos as opções do grau de certeza a duas (*mais certeza* e *menos certeza*). Todo o material utilizado na pesquisa foi disponibilizado num servidor web, de forma que os participantes pudessem fazer o teste remotamente.

¹⁵ docs.google.com

Dados do respondente								
Nome						Idade		
Profissão								
Já usou MercadoLivre antes? (SIM ou NÃO)								
Vendedores a serem analisados								
		Respostas (marcar X somente em uma das quatro)				Tempos		
		Fraudador		Normal				
Nome do vendedor	Data da "foto"	Com mais certeza	Com menos certeza	Com mais certeza	Com menos certeza	Início (hh:mm)	Fim (hh:mm)	Tempo gasto
OI	VIRT	04/08/2008						00:00.0
R	I00	31/07/2008						00:00.0
R	RON	2008 31/07/2008						00:00.0
NI	NGE	RO 02/08/2008						00:00.0
AI	JET2	04/08/2008						00:00.0
C	NIOI	04/08/2008						00:00.0
J	AS20	31/07/2008						00:00.0
SC	2008	31/07/2008						00:00.0
IV	VAR	EL 31/07/2008						00:00.0
D	BRA	04/08/2008						00:00.0
N	I21	31/07/2008						00:00.0
BE	AO	02/08/2008						00:00.0
R	ACI	31/07/2008						00:00.0
BC		02/08/2008						00:00.0
M	ETV	AS 31/07/2008						00:00.0
BC	SUS	31/07/2008						00:00.0
SC	OA	02/08/2008						00:00.0
R	OSI	TS 04/08/2008						00:00.0
G	ART	05/08/2008						00:00.0
P		31/07/2008						00:00.0
S	ENH	31/07/2008						00:00.0
SI	TOT	31/07/2008						00:00.0
SL	..C	02/08/2008						00:00.0
								00:00.0

Figura 8: planilha para anotar respostas (com apelidos truncados)

4.3.3. Instruções aos participantes

Preparamos um texto contendo um resumo de informações diversas sobre atuação fraudulenta nos mercados eletrônicos. Esse texto encontra-se no Anexo 1. Utilizamos como exemplos alguns perfis de fraudadores que encontramos na pesquisa exploratória e, para acelerar o processo de aprendizagem, apontamos alguns indícios de fraude encontrados nos seus anúncios e perfis, aproveitando as descrições sobre estratégias de fraudadores mostradas na seção 2.2.2. Para não influenciar indevidamente os participantes da pesquisa, o autor não analisou os perfis de fraudadores presentes na lista a ser examinada por eles, de modo a evitar a inclusão por inadvertência de indícios específicos destes; apenas foram inspecionadas as qualificações atuais para certificar-se de que havia fortes indícios de que foi cometida a fraude de não-entrega.

Para facilitar a análise das respostas, especialmente quanto ao aspecto temporal, demos as seguintes orientações aos participantes:

- Os vendedores deveriam ser analisados na seqüência dada.
- Uma vez dada uma resposta, ela não deveria ser modificada.

Verificamos o cumprimento dessas duas condições analisando o histórico de alterações das planilhas.

As instruções sofreram pequenas modificações ao longo das primeiras execuções do teste piloto, em função das perguntas e comentários dos participantes. Foram adicionadas respostas a algumas perguntas comuns e o texto foi reorganizado e simplificado para facilitar a leitura.

4.3.4. Amostra utilizada

Foi utilizada uma amostra de conveniência de 26 pessoas, recrutadas pelo autor, das quais 7 já haviam participado em um teste piloto anterior. Posteriormente comparamos o desempenho desses 7 com o dos demais. A Tabela 5 relaciona alguns dados demográficos da amostra.

Tabela 5: dados demográficos da amostra

Faixa etária	Entre 14 a 56 anos, sendo a média 28,3 e a mediana 25,5
Ocupação	14 são profissionais formados nas áreas de informática (5), medicina (3), economia (3), direito (1), engenharia (1), magistério do ensino médio (1) 11 são estudantes universitários 1 é estudante do ensino fundamental
Sexo	4 mulheres e 22 homens
Experiência com o MercadoLivre	13 já tinha utilizado o MercadoLivre antes
Forma de participação	11 fizeram presencialmente (junto do autor) 15 fizeram via Internet

4.4. Resultados

O teste piloto iniciou-se efetivamente no dia 19/08/2008 e terminou no dia 25/08/2008. Os primeiros participantes fizeram-no presencialmente. As dúvidas mais comuns foram incluídas nas instruções e então passamos a aceitar a participação via Internet. Para garantir a validade dos resultados, os participantes foram lembrados verbalmente das regras de execução do teste. A situação atual dos vendedores (quem de fato teve conduta com indícios de fraude) não foi informada a nenhum dos participantes antes do término do teste piloto.

Os resultados das planilhas foram consolidados em um banco de dados e foram incluídas informações adicionais sobre os participantes (dados pessoais, data da execução da tarefa, se esta foi presencial ou não e se já havia participado em um teste piloto anterior).

Na Tabela 6 mostramos um sumário dos resultados, com a quantidade de votos que cada vendedor recebeu apontando-o como fraudador. Também indicamos quais vendedores de fato foram inabilitados e são suspeitos de terem cometido fraude. Observa-se que praticamente todos os vendedores suspeitos estão aglutinados no topo da tabela, com uma grande quantidade de votos. O número do vendedor corresponde à sua posição na tabela fornecida aos participantes. Os vendedores 1, 2 e 3 foram excluídos (cfr. seção 4.3.1).

Tabela 6: resultados da votação

Vendedor	Votos indicando que é fraudador	Fraudador
16	22	SIM
15	21	SIM
7	19	SIM
12	19	SIM
19	17	SIM
21	13	SIM
8	11	Não
9	9	Não
10	8	Não
17	7	Não
18	7	Não
4	7	Não
13	6	Não
11	6	Não
14	5	Não
20	5	Não
6	5	Não
5	5	Não
23	3	Não
22	3	SIM

Com relação às diversas categorias de participantes, fizemos uma comparação na Tabela 7, mostrando o porcentual de participantes que pertenciam a cada categoria e que votaram pela reprovação do vendedor em questão. Estão em negrito os seis primeiros colocados de cada categoria. Em apenas duas delas houve uma mudança significativa na ordenação, avançando um vendedor normal para uma colocação anteriormente ocupada por um fraudador. A inspeção dos dados mostra pouca variação entre as diferentes categorias, sustentando a validade dos resultados.

Tabela 7: comparação das diversas classes de participantes

Vendedor	Tipo	Forma		Conhecimento prévio		Já usou MercadoLivre		Idade		Momento		
		Presencial	Internet	Veteranos	Novos	Sim	Não	> 25,5	< 25,5	< 21/08	> 20/8	
16	F	85%	82%	87%	86%	84%	85%	85%	85%	85%	75%	89%
15	F	81%	91%	73%	100%	74%	77%	85%	77%	85%	88%	78%
12	F	73%	73%	73%	57%	79%	69%	77%	62%	85%	75%	72%
7	F	73%	64%	80%	57%	79%	69%	77%	77%	69%	63%	78%
19	F	65%	64%	67%	43%	74%	69%	62%	46%	85%	50%	72%
21	F	50%	64%	40%	71%	42%	62%	39%	54%	46%	38%	56%
8	N	42%	46%	40%	29%	47%	54%	31%	54%	31%	25%	50%
9	N	35%	36%	33%	0%	47%	39%	31%	31%	39%	25%	39%
10	N	31%	36%	27%	14%	37%	46%	15%	23%	39%	25%	33%
4	N	27%	18%	33%	0%	37%	31%	23%	31%	23%	25%	28%
17	N	27%	9%	40%	14%	32%	23%	31%	31%	23%	25%	28%
18	N	27%	36%	20%	14%	32%	15%	39%	15%	39%	25%	28%
13	N	23%	18%	27%	14%	26%	23%	23%	31%	15%	25%	22%
11	N	23%	9%	33%	0%	32%	31%	15%	23%	23%	13%	28%
20	N	19%	9%	27%	0%	26%	23%	15%	31%	8%	13%	22%
14	N	19%	9%	27%	0%	26%	31%	8%	31%	8%	0%	28%
6	N	19%	9%	27%	0%	26%	15%	23%	31%	8%	0%	28%
5	N	19%	9%	27%	0%	26%	15%	23%	23%	15%	13%	22%
22	F	12%	9%	13%	14%	11%	8%	15%	15%	8%	13%	11%
23	N	12%	18%	7%	0%	16%	8%	15%	8%	15%	0%	17%

Legenda

Veteranos: participaram de um teste piloto anterior
 Momento: dia em que foi respondida a pesquisa
 Células destacadas: seis primeiros colocados

Com relação ao tempo gasto pelos participantes, temos que o tempo total para avaliar os 20 vendedores (excluídos os vendedores 1, 2 e 3) variou de 30 minutos a 2 horas e 41 minutos, com média de 1 hora e 15 minutos. O tempo médio gasto com cada vendedor variou de 2,77 minutos a 4,77 minutos, com média de 3,7 minutos.

O desempenho individual dos participantes está mostrado na Tabela 8, que está em ordem decrescente de $|VE_{FCF}|$ e em ordem crescente de $|VE_{NCF}|$.

Tabela 8: desempenho dos participantes

Participante	$ VE_{FCF} $	FCF	$ VE_{NCF} $	NCF
1	7	100%	2	15%
10	6	86%	2	15%
18	6	86%	4	31%
13	6	86%	6	46%
20	6	86%	7	54%
26	6	86%	7	54%
4	5	71%	1	8%
11	5	71%	3	23%
21	5	71%	5	39%
2	5	71%	5	39%
15	5	71%	6	46%
6	5	71%	7	54%
25	5	71%	7	54%
22	5	71%	8	62%
23	4	57%	0	0%
9	4	57%	0	0%
12	4	57%	0	0%
7	4	57%	0	0%
5	4	57%	1	8%
19	4	57%	6	46%
14	3	43%	1	8%
8	3	43%	2	15%
17	2	29%	0	0%
3	2	29%	0	0%
16	2	29%	2	15%
24	1	14%	2	15%
Média	4,4	63%	3,2	25%
Mediana	5	71%	2	15%
Moda	5	71%	0	0%

Podemos observar pelos dados que os participantes efetivamente foram capazes de discriminar entre vendedores normais e os fraudadores por uma margem

significativa (38% em média). Apenas um dos participantes teve um desempenho muito baixo (o nº 24) e vários tiveram um desempenho muito bom, com $FCF \gg NCF$ (n^{os} 1, 4, 7, 9, 10, 12, 23).

Vamos agora calcular as médias das duas medidas de desempenho para cada um dos grupos diferentes de participantes, incluindo agora uma categoria relacionada como tempo médio para avaliar cada vendedor. Os resultados encontram-se na Tabela 9. A média de FCF é pouco afetada (com exceção do caso dos primeiros participantes – os que fizeram nos dias 19 e 20). Já a média de NCF sofre uma variação maior em dois casos: na categorização por conhecimento prévio e por momento em que a tarefa foi feita. No primeiro caso, notamos que os participantes de um teste piloto anterior foram significativamente mais precisos, com NCF bem menor. Com relação ao momento, percebemos que os primeiros avaliadores classificaram menos vendedores como fraudadores.

Tabela 9: desempenho por categoria

		FCF (média)	NCF (média)
Forma	Presencial	64%	20%
	Internet	62%	28%
Conhecimento prévio	Veteranos	61%	7%
	Novos	63%	32%
Usou MercadoLivre antes	Sim	63%	27%
	Não	63%	22%
Idade	> 25,5	59%	28%
	< 25,5	66%	22%
Momento	19 a 20	57%	16%
	21 a 25	65%	29%
Duração média por vendedor	> 3,7 min	62%	26%
	< 3,7 min	63%	24%

Com relação ao grau de certeza nas respostas, o grau era “maior certeza” em 60,5% das respostas corretas e em 45% das respostas erradas. Ou seja, a insegurança dos participantes de fato estava relacionada com a probabilidade de errar.

4.5. Análise dos resultados

Os resultados obtidos apontam para a viabilidade de utilizar agentes humanos não-especialistas na detecção de fraudadores. No entanto, esse resultado ainda é insuficiente. Como o número de vendedores avaliados é pequeno, a utilização das médias aritméticas de *FCF* e *NCF* como valor esperado das mesmas deve ser feita com cautela, pois o desvio padrão será elevado.

Há ainda possíveis explicações alternativas para os bons resultados obtidos. Uma é a de que alguns dos fraudadores presentes na lista do teste piloto sejam na verdade *os mesmos* que os apresentados nos exemplos, só que com outros nomes. Nesse caso, os indícios fornecidos no material de treinamento introduziriam um viés.

Outra explicação alternativa é a de que o grupo de fraudadores utilizado não seja representativo: podem ser mais fáceis de detectar que outros. Seria interessante repetir o teste piloto com outro grupo de fraudadores, de preferência escolhido aleatoriamente dentro de um universo maior.

O não-seguimento das regras do estudo, especialmente no caso dos participantes que fizeram a tarefa via Internet, também poderia comprometer a validade. No entanto, não obtivemos nenhum indício de que isso tenha ocorrido.

A quantidade de fraudadores presentes na amostra não é representativa do que acontece nos mercados eletrônicos, mesmo utilizando algum tipo de pré-seleção, que no nosso caso gerou um conjunto com 10% de fraudadores, inferior aos 30% presentes na amostra (incluídos os três primeiros para teste). Apesar dos participantes não terem sido informados acerca dessa quantidade e de terem que analisar seis vendedores normais antes do primeiro fraudador, não podemos afirmar com segurança que os resultados seriam os mesmos, já que seria necessário analisar muito mais vendedores para achar um fraudador. Isso poderia tornar outros fatores relevantes, e.g. o cansaço.

Por fim, a amostra utilizada, apesar de englobar pessoas com características demográficas variadas, não é aleatória, o que compromete a generalização dos resultados. No entanto, é significativo o fato de que os resultados obtidos por pessoas de diferentes categorias variaram pouco.