



Paulo Alexandre Rua Trindade

**Modelos de medição psicométrica:
avaliação das valências matemáticas
dos alunos do 1.^o ano da FCUP à
entrada do ensino superior**

Departamento de Matemática
Faculdade de Ciências da Universidade do Porto
2015

Paulo Alexandre Rua Trindade

**Modelos de medição psicométrica:
avaliação das valências matemáticas
dos alunos do 1.^o ano da FCUP à
entrada do ensino superior**

*Relatório de Estágio submetido à Faculdade de Ciências da
Universidade do Porto para obtenção do grau de Mestre
em Engenharia Matemática*

Departamento de Matemática
Faculdade de Ciências da Universidade do Porto
2015

Para Amélia e Jorge, Isabel e Patrícia

Agradecimentos

No final desta etapa da minha vida não posso deixar de agradecer aos que ao longo do presente percurso académico me ajudaram a caminhar.

Começo por agradecer, de modo genérico, à Faculdade de Ciências da Universidade do Porto, pela confiança que depositou na minha pessoa ao partilhar o trabalho que foi desenvolvido no âmbito dos testes diagnósticos de Matemática. De igual modo, expresso o meu agradecimento à Faculdade de Medicina da Universidade do Porto por ter permitido a realização deste estágio, no qual tive a oportunidade de aliar o desenvolvimento dos meus conhecimentos académicos com a minha experiência enquanto professor de Matemática do 3.^o ciclo do Ensino Básico e do Ensino Secundário.

Não posso também deixar de manifestar o meu apreço e gratidão, de modo particular, a algumas das pessoas que contribuíram para que este trabalho fosse possível.

À minha orientadora da Faculdade de Ciências da Universidade do Porto, Prof Ana Rita Gaio, e ao meu orientador da Faculdade de Medicina da Universidade do Porto, Prof Milton Severo, pela sabedoria, partilha de conhecimento, dedicação, empenho, disponibilidade, atenção dispensada, opiniões, críticas, constante incentivo e inestimável contributo para a realização deste trabalho.

À Prof Paula Lomp, do Departamento de Matemática, ao Prof José Paulo Leal e à Prof Ana Paula Tomás, do Departamento de Ciência de Computadores, pela disponibilidade que demonstraram para facultar os dados necessários para a realização deste trabalho e para esclarecer questões relacionadas com os mesmos.

Aos meus pais, Amélia e Jorge, pelo apoio e constante preocupação... tenho a felicidade de ter os pais com que qualquer filho sonha. À Isabel pelo ânimo, presença e compreensão e por ter sido a pessoa que mais me incentivou a dar o passo mais importante desta caminhada... o primeiro. À minha irmã Patrícia por ser uma fonte de inspiração e de orgulho pela fantástica pessoa que é.

Aos meus restantes familiares, amigos, colegas de trabalho, colegas e professores do mestrado pela sua paciência, disponibilidade e recomendações que me foram dando ao longo deste trabalho.

A todos o meu profundo e sentido obrigado.

Resumo

Um número significativo de alunos que ingressam no ensino superior, em particular nas licenciaturas com uma forte componente matemática, revelam grandes lacunas e cometem erros graves em operações elementares de aritmética ou cálculo algébrico. Demonstram ainda uma grande inaptidão para compreenderem algoritmos e/ou raciocínios lógico-formais.

Para fazer face a este problema, cuja gravidade se tem vindo a acentuar nos anos mais recentes, a Faculdade de Ciências da Universidade do Porto tem vindo a realizar testes diagnósticos de Matemática. O teste não é obrigatório e destina-se aos alunos do primeiro ano.

Aos alunos que obtiverem classificações baixas no teste diagnóstico é disponibilizada uma plataforma online de treino, com o objetivo de colmatar as lacunas evidenciadas.

Até 2010 os testes diagnósticos eram da responsabilidade do Departamento de Matemática e foram distribuídos aos alunos em papel. Após esta data, e com a colaboração do Departamento de Ciência de Computadores, os testes passaram a ser realizados no computador. Além disso, foi também disponibilizada uma plataforma online, constituída por um banco de itens semelhantes aos presentes nos testes diagnósticos, à qual os alunos podem aceder para treinar e melhorar as suas aptidões matemáticas e, consequentemente, obter bons resultados na avaliação da unidade curricular de Cálculo.

Este trabalho teve como principais objetivos a análise da qualidade dos testes aplicados, a avaliação da tendência temporal dos resultados obtidos e o estudo do impacto do treino na plataforma nos resultados obtidos pelos alunos. Para o efeito, foi usada a Teoria Clássica de Testes, modelos da Teoria de Resposta ao Item e técnicas de calibração.

Os resultados mostraram que os testes diagnósticos revelaram boa qualidade e apresentaram um nível de fiabilidade bastante satisfatório. Observou-se também que a esmagadora maioria dos itens presentes nos testes eram discriminativos.

Os conteúdos matemáticos em que os estudantes revelaram maior dificuldade foram: as inequações envolvendo funções irracionais, logarítmicas e exponenciais; o cálculo de derivadas de funções irracionais e exponenciais e o cálculo de limites com logaritmos. Destacaram-se entre os conteúdos matemáticos mais acessíveis para os discentes: as operações com números racionais e as equações do primeiro grau.

Verificou-se existirem diferenças estatisticamente significativas entre as habilidades médias das diferentes coortes de entrada. Em média o grupo de estudantes que revelou maior habilidade Matemática foi o que realizou o teste diagnóstico aplicado no ano 2009. O grupo de estudantes com menor habilidade média foi o que realizou o teste diagnóstico em 2011.

A plataforma online de treino, disponibilizada aos alunos para que pudessem colmatar as suas dificuldades, revelou servir os propósitos para que foi criada. O treino teve impacto positivo, estatisticamente significativo, nos resultados obtidos pelos discentes.

Abstract

A significant number of undergraduate students, particularly those enrolling courses with a strong mathematical component, reveal major gaps and commit serious errors in elementary operations of arithmetic or algebraic calculations. Demonstrate also a great inability to understand algorithms and/or logical-formal reasoning.

To address this problem, whose severity has been increasing over the last years, a Mathematics diagnostic test has been performed at the Faculty of Sciences of the University of Porto. The test is non-compulsory and has been directed to first-year students.

The students that score low values at the diagnostic test are advised to attend the Pre-Calculus course, with the aim to improve the gaps highlighted above.

Until 2010 the diagnostic tests were under the responsibility of the Mathematics Department and were distributed to the students in paper. From there onwards, and with the collaboration of the Computer Science Department, the tests have been carried out online. In addition, an online platform was also provided, consisting of a bank of items similar to those present in the diagnostic tests, which students can access to train and improve their mathematical skills and therefore obtain good results in the evaluation of the course of Calculus.

This work had as main objectives the analysis of the quality of the tests, the evaluation of the time trend of the results and the analysis of the impact of the study on the platform on the results obtained by the students. Classical test theory, item response theory models and calibration techniques were used.

The results showed that the tests had a good quality and had a very satisfactory level of reliability. It was also observed that the majority of the items in the tests were discriminatory.

The mathematical contents in which students showed greater difficulty were: inequalities involving irrational, logarithmic and exponential functions; derivative of a irrational or exponential function and limits with logarithms. Stood out among the most affordable mathematical contents: operations with rational numbers and first degree equations.

There were statistically significant differences between the average abilities of different inlet cohorts. On average the group of students who showed greater Math ability was what held the diagnostic test applied in 2009. The group of students with lower average ability was the one who held the 2011 diagnostic test.

The online training platform, available to students so that they could overcome their difficulties, revealed to serve the purposes for which it was created. The training had a statistically significant positive impact on the results obtained by the students.

”Nem tudo o que pode ser contado conta,
e nem tudo o que conta pode ser contado.”

Albert Einstein

Conteúdo

Resumo	5
Abstract	6
Índice de Tabelas	11
Índice de Figuras	13
1 Introdução	14
2 Métodos	17
2.1 Teoria Clássica de Testes (TCT)	17
2.2 Teoria de Resposta ao Item (TRI)	20
2.3 Comparação entre Teoria Clássica de Testes e Teoria de Resposta ao Item . .	27
2.4 Calibração	28
2.4.1 Calibração usando Teoria Clássica de Testes	28
2.4.2 Calibração usando Teoria de Resposta ao Item	30
3 Aplicação das metodologias aos dados dos testes	33
3.1 Ano 2009	33
3.1.1 Descrição dos dados/Análise descritiva e exploratória	33
3.1.2 Resultados	33
3.1.2.1 Teoria Clássica de Testes (TCT)	35
3.1.2.2 Teoria de Resposta ao Item (TRI)	36
3.2 Ano 2010	44
3.2.1 Descrição dos dados/Análise descritiva e exploratória	44
3.2.2 Resultados	44
3.2.2.1 Teoria Clássica de Testes (TCT)	45
3.2.2.2 Teoria de Resposta ao Item (TRI)	46
3.3 Ano 2011	53
3.3.1 Descrição dos dados/Análise descritiva e exploratória	53
3.3.2 Resultados	53
3.3.2.1 Teoria Clássica de Testes (TCT)	55
3.3.2.2 Teoria de Resposta ao Item (TRI)	57
3.4 Ano 2012	63
3.4.1 Descrição dos dados/Análise descritiva e exploratória	63
3.4.2 Resultados	63

3.4.2.1	Teoria Clássica de Testes (TCT)	64
3.4.2.2	Teoria de Resposta ao Item (TRI)	66
3.5	Ano 2013	72
3.5.1	Descrição dos dados/Análise descritiva e exploratória	72
3.5.2	Resultados	72
3.5.2.1	Teoria Clássica de Testes (TCT)	73
3.5.2.2	Teoria de Resposta ao Item (TRI)	75
3.6	Síntese de resultados	81
3.7	Calibração	82
3.7.1	Calibração usando Teoria Clássica de Testes	83
3.7.2	Calibração usando Teoria de Resposta ao Item	88
4	Avaliação do impacto do treino	93
4.1	Descrição do problema	93
4.2	Descrição dos dados	93
4.3	Descrição estatística dos dados	93
4.3.1	Variável dependente: Avaliacao	93
4.3.2	Variáveis explicativas	94
4.3.2.1	Curso	94
4.3.2.2	TD	96
4.3.2.3	Tempo	96
4.4	Ajustamento do modelo	97
4.5	Diagnóstico e estudo do modelo final	100
4.5.1	<i>Outliers</i>	100
4.5.2	<i>Leverages</i>	100
4.5.3	Observações influentes	100
4.5.4	Homocedasticidade	102
4.5.5	Normalidade dos resíduos estandardizados	104
4.6	Interações	104
5	Conclusões	106
A	Teste Diagnóstico 2010	112

Lista de Tabelas

1	Organização dos itens (2009)	34
2	Parâmetros estimados usando Teoria Clássica de Testes e Teoria de Resposta ao Item - Modelo 2PL (2009)	37
3	Percentis do χ^2 - Modelo 1PL (2009)	37
4	Percentis do χ^2 - Modelo 2PL (2009)	37
5	Percentis das razões de verossimilhança (2009)	39
6	Avaliação marginal da qualidade do ajustamento - Modelo 1PL (2009) . . .	39
7	Avaliação marginal da qualidade do ajustamento - Modelo 2PL (2009) . . .	40
8	Respostas aos itens 15, 26 e 28 por opção (A, B, C ou D), para o ano de 2009. As células a verde representam a opção correta.	41
9	Parâmetros estimados usando Teoria Clássica de Testes e Teoria de Resposta ao Item - Modelo 2PL (2010)	46
10	Percentis do χ^2 - Modelo 1PL (2010)	47
11	Percentis do χ^2 - Modelo 2PL (2010)	47
12	Percentis das razões de verossimilhança (2010)	48
13	Avaliação marginal da qualidade do ajustamento - Modelo 1PL (2010) . . .	48
14	Respostas ao item 15 por opção (A, B, C ou D), para o ano de 2010. A célula a verde representa a opção correta.	50
15	Organização dos itens (2011)	54
16	Parâmetros estimados usando Teoria Clássica de Testes e Teoria de Resposta ao Item - Modelo 2PL (2011)	56
17	Percentis do χ^2 - Modelo 1PL (2011)	57
18	Percentis do χ^2 - Modelo 2PL (2011)	58
19	Percentis das razões de verossimilhança (2011)	59
20	Parâmetros estimados usando Teoria Clássica de Testes e Teoria de Resposta ao Item - Modelo 2PL (2012)	65
21	Percentis do χ^2 - Modelo 1PL (2012)	66
22	Percentis do χ^2 - Modelo 2PL (2012)	67
23	Percentis das razões de verossimilhança (2012)	67
24	Avaliação marginal da qualidade do ajustamento - Modelo 1PL (2012) . . .	68
25	Avaliação marginal da qualidade do ajustamento - Modelo 2PL (2012) . . .	68
26	Parâmetros estimados usando Teoria Clássica de Testes e Teoria de Resposta ao Item - Modelo 2PL (2013)	74
27	Percentis do χ^2 - Modelo 1PL (2013)	75
28	Percentis do χ^2 - Modelo 2PL (2013)	76
29	Percentis das razões de verossimilhança (2013)	76

30	Total de perguntas por teste diagnóstico (diagonal) e número de perguntas âncora	81
31	Análise da qualidade dos itens de acordo com a Teoria Clássica de Testes . .	81
32	Análise da qualidade dos itens de acordo com o modelo 2PL da Teoria de Resposta ao Item	81
33	Estatísticas das classificações observadas	83
34	Estatísticas das habilidades estimadas usando TCT, referência 2009	86
35	Classificações calibradas tendo como referência o ano 2009	87
36	Constantes de ligação referência 2009	88
37	Constantes de calibração de acordo com a TCT e a TRI	88
38	Estatísticas das habilidades estimadas usando TRI, referência 2009	88
39	Classificação verdadeira dado θ	90
40	Classificação verdadeira (escala 0-20) dado θ	91
41	Estatísticas numéricas descritivas da variável Avaliacao em função da variável Curso	94
42	Estatísticas numéricas da variável Avaliacao em função da variável Treino . .	97
43	Modelo 1	98
44	ANOVA Modelo 1	98
45	VIF Modelo 1	99
46	Modelo 2	99
47	ANOVA Modelo 2	99

Lista de Figuras

1	Curvas características dos itens (Modelo 1PL)	24
2	Curva de informação total	25
3	Erro padrão de medição	25
4	Classificações obtidas (2009)	35
5	Percentagem de respostas corretas - para os itens (2009)	36
6	Estatísticas de teste do χ^2 - Modelo 1PL (2009)	38
7	Estatísticas de teste do χ^2 - Modelo 2PL (2009)	38
8	Estatísticas de teste da razão de verossimilhanças (2009)	39
9	Habilidades previstas pelo modelo da TRI (2009)	40
10	Curvas características dos itens (2009)	42
11	Curva de informação total (2009)	42
12	Erro padrão de medição (2009)	43
13	Valores próprios e componentes principais (2009)	43
14	Classificações obtidas (2010)	44
15	Percentagem de respostas corretas - para os itens (2010)	45
16	Estatísticas de teste do χ^2 - Modelo 1PL (2010)	47
17	Estatísticas de teste do χ^2 - Modelo 2PL (2010)	48
18	Estatísticas de teste da razão de verossimilhanças (2010)	49
19	Habilidades previstas pelo modelo da TRI (2010)	50
20	Curvas características dos itens (2010)	50
21	Curva de informação total (2010)	51
22	Erro padrão de medição (2010)	52
23	Valores próprios e componentes principais (2010)	52
24	Classificações obtidas (2011)	55
25	Percentagem de respostas corretas - para os itens (2011)	57
26	Estatísticas de teste do χ^2 - Modelo 1PL (2011)	58
27	Estatísticas de teste do χ^2 - Modelo 2PL (2011)	58
28	Estatísticas de teste da razão de verossimilhanças (2011)	59
29	Habilidades previstas pelo modelo da TRI (2011)	60
30	Curvas características dos itens (2011)	61
31	Curva de informação total (2011)	61
32	Erro padrão de medição (2011)	62
33	Valores próprios e componentes principais (2011)	62
34	Classificações obtidas (2012)	63
35	Percentagem de respostas corretas - para os itens (2012)	64
36	Estatísticas de teste do χ^2 - Modelo 1PL (2012)	66

37	Estatísticas de teste do χ^2 - Modelo 2PL (2012)	67
38	Estatísticas de teste da razão de verossimilhanças (2012)	68
39	Habilidades previstas pelo modelo da TRI (2012)	69
40	Curvas características dos itens (2012)	70
41	Curva de informação total (2012)	70
42	Erro padrão de medição (2012)	71
43	Valores próprios e componentes principais (2012)	71
44	Classificações obtidas (2013)	72
45	Percentagem de respostas corretas - para os itens (2013)	73
46	Estatísticas de teste do χ^2 - Modelo 1PL (2013)	75
47	Estatísticas de teste do χ^2 - Modelo 2PL (2013)	76
48	Estatísticas de teste da razão de verossimilhanças (2013)	77
49	Habilidades previstas pelo modelo da TRI (2013)	78
50	Curvas características dos itens (2013)	78
51	Curva de informação total (2013)	79
52	Erro padrão de medição (2013)	80
53	Valores próprios e componentes principais (2013)	80
54	Percentagem média de acerto dos estudantes por ano em todas as perguntas do teste e nas perguntas âncora	82
55	Distribuições das classificações nos testes e nos itens âncora (2009)	83
56	Distribuições das classificações nos testes e nos itens âncora (2010)	84
57	Distribuições das classificações nos testes e nos itens âncora (2011)	84
58	Distribuições das classificações nos testes e nos itens âncora (2012)	85
59	Distribuições das classificações nos testes e nos itens âncora (2013)	85
60	Evolução temporal da habilidade média estimada usando TCT	86
61	Evolução temporal da habilidade média estimada usando TRI	89
62	Comparação da evolução temporal da habilidade média estimada usando TCT e TRI	92
63	Representações gráficas da variável dependente: Avaliacao	94
64	Representações gráficas da variável explicativa Curso	95
65	Representações gráficas da variável explicativa TD	96
66	Representações gráficas da variável Avaliacao em função da variável Treino	97
67	Resíduos (Modelo 1)	98
68	Resíduos (Modelo 2)	100
69	<i>Leverages</i> (Modelo 2)	101
70	Distâncias de Cook (Modelo 2)	101
71	<i>Leverages</i> / Distâncias de Cook (Modelo 2)	102
72	Seminormal (Modelo 2)	102
73	Valores ajustados / Resíduos estandardizados (Modelo 2)	103
74	Valores ajustados / Resposta (Modelo 2)	103
75	Análise da normalidade dos resíduos estandardizados (Modelo 2)	104
76	Interação entre as variáveis TD e Treino	105
77	Interação entre as variáveis TD e Curso	105

Capítulo 1

Introdução

Um número significativo de alunos que ingressam no ensino superior, em particular nas licenciaturas com uma forte componente matemática, revelam grandes lacunas e cometem erros graves em operações elementares de aritmética ou cálculo algébrico, para além de demonstrarem uma grande inaptidão para compreenderem algoritmos e/ou raciocínios lógico-formais. Esta diferença, entre o nível de preparação esperado ou exigido no momento da entrada no ensino superior e as capacidades e conhecimentos matemáticos adquiridos no ensino secundário, tornou-se conhecida como "o problema da matemática" [46].

Nos últimos anos, governos, empregadores e instituições de ensino superior de vários países manifestaram a sua preocupação face a este problema. Foram também elaborados vários artigos e relatórios que realçaram esta preocupação [26], [39], [46], [49], .

Estudos recentes sobre esta problemática levaram muitas instituições de ensino superior a organizar alguma forma de prestar apoio matemático aos seus alunos, com o objetivo principal de os ajudar a superar as suas dificuldades [19], e a usar testes diagnósticos de Matemática como forma de aferir as capacidades e conhecimentos matemáticos dos seus novos alunos, em particular dos que frequentam cursos com forte componente Matemática. Na maior parte das instituições os testes diagnósticos são constituídos por questões de escolha múltipla, agrupadas por temas, e realizados em papel ou no computador [35].

Vários estudos internacionais mostram que o apoio prestado por instituições de ensino superior (Irlanda, Reino Unido, Holanda, Austrália) tem efeitos positivos no percurso académico dos estudantes [14], [19], [20], [29], [36], [39], [47], [56]. Em Portugal, o Instituto Superior Técnico levou a efeito no início do ano letivo 2002/2003 um teste diagnóstico de Matemática, dirigido aos alunos do primeiro ano das suas licenciaturas. O processo identificou várias lacunas, inclusivamente, em matérias anteriores ao Ensino Secundário tais como simplificação de frações ou regras de potenciação [45].

Para fazer face a este problema, cuja gravidade parece ter-se vindo a acentuar nos anos mais recentes, foi decidido na reunião 01/2009 da Comissão Científica Alargada do Departamento de Matemática Pura da FCUP realizar uma prova de aferição de conhecimentos fundamentais de Matemática (teste diagnóstico) para os alunos que ingressam na 1.^a fase das licenciaturas de Matemática e, eventualmente, noutras licenciaturas da FCUP. O principal objetivo deste teste diagnóstico é identificar carências de carácter elementar na formação dos alunos ao longo do 3.^o Ciclo do Ensino Básico e do Ensino Secundário, ou seja, entre o 7.^o e o 12.^o ano, inclusive, não tendo qualquer consequência na sua posterior avaliação nas diversas disciplinas. O teste enquadrou-se num conjunto de iniciativas que pretendiam

ajudar o aluno a iniciar o estudo nas disciplinas de Matemática das várias licenciaturas, de forma rápida e bem sucedida. Em particular, os resultados do teste diagnóstico sugeriram a frequência do curso de pré-cálculo e o acesso a uma plataforma online de treino, para que os alunos pudessem superar eventuais carências de formação, e para outras iniciativas (site de apoio em Matemática elementar, <http://cmup.fc.up.pt/cmup/apoiomat/>) que ajudassem a uma melhoria substancial do desempenho dos alunos.

Os testes consistiram de 30 perguntas de escolha múltipla, cada uma com quatro opções. Por opção das várias equipas responsáveis pela aplicação dos testes diagnósticos não houve lugar a descontos por respostas erradas. Para simplificação da análise, o teste foi considerado como sendo constituído apenas por itens binários, isto é, itens para os quais se admitem apenas duas respostas possíveis: certa ou errada. Na primeira situação o aluno obtém a cotação de 1 ponto, na segunda de 0 pontos. O facto de não haver descontos implica que as ausências de resposta aos itens sejam tratadas como respostas incorretas. A "classificação final do teste" corresponde à soma das cotações correspondentes à totalidade dos itens.

O trabalho desenvolvido neste estágio teve como principais objetivos analisar a qualidade de todos os testes diagnósticos de Matemática que foram realizados na FCUP ao longo dos últimos anos, em concreto de 2009 a 2013, a modelação e calibração dos resultados e ainda a avaliação do impacto do treino na plataforma no desempenho dos alunos na unidade curricular de Cálculo. Nesta sequência, espera-se, desta forma, poder contribuir para um melhor entendimento da progressão temporal das valências matemáticas dos alunos do primeiro ano da FCUP, à entrada no ensino superior, e permitir, em particular, fornecer uma análise crítica ao trabalho que tem vindo a ser realizado na FCUP, nomeadamente, no que se refere à qualidade dos testes diagnósticos, com sugestão de possíveis alterações, e avaliar a importância da plataforma online de treino na colmatação das lacunas evidenciadas pelos estudantes.

Nesta análise recorreu-se à Teoria Clássica de Testes (TCT) e à Teoria de Resposta ao Item (TRI). Ambas as teorias consideram o teste constituído por itens binários, isto é, itens para os quais apenas existe a possibilidade de se obter uma resposta certa ou errada.

A Teoria Clássica de Testes surgiu no início do século XX com o desenvolvimento da Psicometria e assentou, principalmente, em três factos: reconhecimento da presença de erros nas medições; conceção do erro como variável aleatória e conceito de correlação. Os trabalhos realizados por Charles Spearman, em 1904, relacionados com o coeficiente de correlação e com o erro de medição podem ser considerados o início da Teoria Clássica de Testes [31], [42], [58]. O principal fundamento da Teoria Clássica de Testes é que a classificação observada de um indivíduo num dado teste, representada por X , pode ser decomposta na classificação verdadeira do indivíduo, designada por T , e num erro aleatório, designado por E , da seguinte forma:

$$X = T + E.$$

A classificação verdadeira de um indivíduo, T , é definida como o valor esperado da classificação observada se o mesmo teste for aplicado ao indivíduo um número infinito de vezes [12].

A Teoria de Resposta ao Item, geralmente aceite como um desenvolvimento da Teoria Clássica de Testes, surge no início da década de 50 do século XX com os trabalhos desenvolvidos pelo psicometrista Frederic Lord, pelo matemático Georg Rasch e pelo sociologista Paul

Lazarsfeld. Esta teoria assenta basicamente em três pressupostos [25]: um traço latente unidimensional, geralmente denotado por θ , e designado por habilidade; a existência de independência local dos itens condicionada pela habilidade e a assunção de que a resposta de um indivíduo a um item pode ser modelada por uma função logística em função das características do item e da habilidade do indivíduo no traço que se pretende avaliar. A utilização da Teoria de Resposta ao Item apenas se generalizou nas décadas de 70 e 80, por um lado, devido à massificação dos computadores pessoais que permitiram aos investigadores usufruir da sua capacidade de cálculo e, por outro, quando a sua utilidade e vantagens face à Teoria Clássica de Testes se tornaram evidentes [8], [25].

Na Teoria Clássica de Testes o objeto de análise é o teste no seu todo ao passo que na Teoria de Resposta ao Item o objeto de estudo é cada um dos itens presentes no teste. Nesta última, as classificações obtidas pelos examinandos são independentes do teste efetuado e as características dos itens são independentes do grupo de examinandos, o que não sucede na Teoria Clássica de Testes.

O nível de significância foi fixado em 0,05. Todas as análises estatísticas foram realizadas com o software "R Project for Statistical Computing", versões 2.13.2 e 3.0.1 (R Foundation, Viena, Áustria), mais especificamente, recorreu-se, entre outros, aos pacotes: "ltm", para efetuar análises relacionadas com o ajustamento de modelos de traço latente; "equate" para calibração dos resultados usando a Teoria Clássica de Testes; "plink" para calibração dos resultados usando a Teoria de Resposta ao Item; "stats" e "nlme" para análise do impacto do treino por ajustamento de modelos lineares.

Este relatório de estágio encontra-se estruturado da seguinte forma: introdução, métodos, aplicação dos métodos aos dados dos testes, avaliação do impacto do treino e conclusões. No capítulo "Métodos" são apresentadas as metodologias seguidas, nomeadamente, a Teoria Clássica de Testes, a Teoria de Resposta ao Item, uma comparação entre estas duas teorias, com as vantagens e desvantagens de cada uma delas, e diferentes processos de calibração de habilidades. No capítulo "Aplicação das metodologias aos dados dos testes" são aplicados os métodos aos dados dos testes diagnósticos da FCUP dos anos 2009 a 2013 sendo descritos os dados e apresentados os resultados obtidos. No capítulo "Avaliação do impacto do treino" é estudado o impacto do treino dos alunos na plataforma sobre os resultados obtidos na unidade curricular de Cálculo por ajustamento de modelos de regressão linear múltipla. No capítulo "Conclusões" são apresentadas as principais conclusões resultantes do estudo efetuado.

Capítulo 2

Métodos

2.1 Teoria Clássica de Testes (TCT)

A bibliografia correspondente a este capítulo é essencialmente DeChamplain (2010) [12] e DeVellis (2006) [13].

Segundo DeChamplain, a classificação num teste é um número que supostamente reflete a habilidade de um indivíduo numa determinada área de conhecimento claramente definida. O facto de não existir nenhum teste perfeito e de todas as classificações num teste terem associado um erro de medição, não correlacionado com a área de conhecimento avaliada, faz com que seja necessário um modelo da teoria de testes que ajude a compreender melhor: a relação existente entre a classificação observada num teste e a habilidade subjacente no domínio que, geralmente, é não observada; o processo de medição e como este é afetado por fontes de erro [13].

A Teoria Clássica de Testes consiste num conjunto de conceitos e técnicas matemáticas que pretendem servir de base à conceção e análise de vários instrumentos de medida, funcionando naturalmente como ponto de referência para abordagens recentes de técnicas de medição.

Muitos fenómenos relacionados sobretudo com o ensino e com a saúde, tais como o conhecimento de um aluno numa determinada matéria ou o grau de sofrimento de um paciente não são diretamente observáveis. O que se pretende é encontrar indicadores que forneçam informação precisa sobre essas variáveis não observáveis. A Teoria Clássica de Testes compreende um conjunto de princípios que permitem determinar se os indicadores fornecem boas estimativas das variáveis não observadas. Essencialmente, consiste em utilizar informação observável, por exemplo, a classificação observada num teste ou as respostas a um inquérito, para conhecer variáveis que não são diretamente observáveis.

Inferir sobre algo que não é diretamente observável é um processo naturalmente associado a um determinado erro. Charles Spearman [50] foi um dos primeiros matemáticos a reconhecer e a explicar esta relação entre a informação recolhida e a informação que verdadeiramente interessa, a variável não observável. A diferença entre estas duas informações corresponde ao chamado erro de medição.

Vários testes, de cariz académico ou não, medem variáveis dos indivíduos que não são diretamente observáveis, tais como a inteligência ou a motivação para uma determinada tarefa. Além disso, a classificação observada de um indivíduo num teste resulta sobretudo

da sua habilidade na área de conhecimentos avaliada no teste. Contudo esta classificação pode também ser influenciada por outros fatores, por exemplo, ambiguidade dos itens, pouco interesse em obter uma boa classificação, problemas de saúde, distração ou cansaço do indivíduo. Devido a estes fatores a classificação observada pode não ser uma avaliação precisa da verdadeira habilidade do indivíduo.

A classificação verdadeira de um indivíduo num teste corresponde à classificação observada sem a influência de efeitos aleatórios tais como os acima mencionados. Note-se que os erros sistemáticos resultantes de características permanentes do indivíduo, por exemplo, deficiente capacidade de leitura, são considerados como fazendo parte da classificação verdadeira.

O erro de medição é a diferença entre a classificação observada e a classificação verdadeira e corresponde aos fatores aleatórios que influenciam a classificação observada.

É, portanto, possível escrever, para cada item i , a classificação observada (X_i) como uma soma de duas componentes, a classificação verdadeira (T_i) e o erro de medição (E_i),

$$X_i = T_i + E_i. \quad (1)$$

Definindo $T = \sum_i T_i$, $X = \sum_i X_i$ e $E = \sum_i E_i$, tem-se obviamente:

$$X = T + E. \quad (2)$$

Se se considerar que os erros são aleatórios, não correlacionados (para os vários itens) e não correlacionados com as classificações verdadeiras, por (2) tem-se que a variância das classificações obtidas é igual à soma da variância das classificações verdadeiras com a variância dos erros de medição. Sendo S_X^2 a variância das classificações obtidas, S_T^2 a variância das classificações verdadeiras e S_E^2 a variância dos erros de medição tem-se

$$S_X^2 = S_T^2 + S_E^2. \quad (3)$$

A demonstração desta igualdade é trivial no caso unidimensional e pode ser encontrada em Gulliksen (1950), no caso geral [21].

Para a maior parte dos instrumentos de medição encontram-se definidas, essencialmente, duas propriedades psicométricas: a fiabilidade e a validade.

De um modo geral, a validade de um instrumento pode ser definida como a sua capacidade de realmente medir aquilo que se propõe a medir [42]. Mais especificamente, refere-se à capacidade do instrumento de efetivamente medir um conceito teórico específico, o construto, seja este um processo psicológico ou uma característica dos indivíduos [53], [55].

A fiabilidade de um teste, que também pode ser designada por consistência interna ou precisão, diz respeito à reprodutibilidade da medida, ou seja, ao grau de concordância definido pelas múltiplas medidas de um sujeito entre si, e entre medidas de outro sujeito [2]. As medições devem ser fiáveis, no sentido em que os examinadores possam confiar na precisão do instrumento de medida utilizado, sendo a fiabilidade definida como o quociente entre a variância da classificação verdadeira e a variância total da medição [55].

Os erros sistemáticos não afetam a fiabilidade do teste, mas podem afetar negativamente a sua validade.

A fiabilidade de um teste, designada por $r_{XX'}$, é definida como o quociente entre a variância da classificação verdadeira e a variância da classificação obtida:

$$r_{XX'} = \frac{S_T^2}{S_X^2}. \quad (4)$$

A fiabilidade indica a proporção da variância das classificações observadas que se pode atribuir à variância das classificações verdadeiras [12], ou seja, às diferenças na habilidade dos indivíduos e não a outros fatores e pode ser também escrita da seguinte forma:

$$r_{XX'} = 1 - \frac{S_E^2}{S_X^2}. \quad (5)$$

As classificações de um teste têm que ser fiáveis, ou seja, a fiabilidade tem que estar próxima de 1, para que os examinadores possam retirar conclusões válidas sobre as mesmas. A fiabilidade não pode ser estimada diretamente, uma vez que a classificação verdadeira não é conhecida e portanto S_T^2 é desconhecida.

A Teoria Clássica de Testes fornece uma solução para este problema. Considerem-se dois itens estritamente paralelos, ou seja, dois itens que verificam as seguintes condições:

- 1) A variável latente é unidimensional, ou seja, os itens estão relacionados com a mesma classificação verdadeira;
- 2) O erro associado a cada um dos itens é independente da classificação verdadeira e dos erros que surgem dos restantes itens;
- 3) A covariância dos itens com a classificação verdadeira é igual para todos os itens;
- 4) A proporção da contribuição do erro associado a cada item para a variância total é a mesma para todos os itens.

Sobre estas condições de paralelismo estrito, a correlação entre dois quaisquer itens é igual ao produto das correlações de cada um dos itens com a classificação verdadeira [13].

Designando os itens por X e Y tem-se:

$$r_{XY} = r_{XT} \times r_{YT} = r_{XT}^2 = r_{YT}^2. \quad (6)$$

Esta igualdade permite determinar uma estimativa da fiabilidade dos itens.

As condições de paralelismo estrito são bastante exigentes. Contudo, foram desenvolvidos modelos com condições ligeiramente menos exigentes que permitem estimar a fiabilidade de cada item a partir da correlação entre dois dos itens [13].

Na prática, recorre-se ao coeficiente alfa de Cronbach (α) que mede a consistência interna de um teste e estima a capacidade de discriminação do teste, em relação aos indivíduos [10]. O coeficiente alfa de Cronbach é dado pela seguinte expressão:

$$\alpha = \frac{k}{k-1} \left(1 - \frac{\sum_{i=1}^k \sigma_i^2}{\sigma_X^2} \right) \quad (7)$$

sendo k o número de itens, σ_i^2 a variância do item (binário) i e σ_X^2 a variância da classificação observada no teste [11].

Se os itens forem classificados dicotomicamente, pode-se usar a distribuição de Bernoulli para cada item e a expressão (7) pode ser reescrita da seguinte forma:

$$\alpha = \frac{k}{k-1} \left(1 - \frac{\sum_{i=1}^k p_i(1-p_i)}{\sigma_X^2} \right) \quad (8)$$

sendo p_i a percentagem de respostas corretas ao item i .

O erro padrão de medição (SEM, Standard Error of Measurement) é o desvio padrão dos erros de medição associados às classificações de um teste para um grupo específico de indivíduos. Partindo da equação (5) tem-se

$$S_E^2 = S_X^2(1 - r_{XX'}) \quad (9)$$

consequentemente,

$$SEM = S_X \sqrt{1 - r_{XX'}}. \quad (10)$$

O erro padrão de medição permite fazer afirmações sobre a precisão das classificações de cada indivíduo no teste, encontrando-se na mesma unidade que as classificações do teste. Em particular, o erro de medição permite obter intervalos de confiança para as classificações finais.

Os itens mais correlacionados uns com os outros são mais correlacionados com a verdadeira classificação da variável não observável em estudo e são, por isso, melhores itens. Diz-se que estes itens são mais discriminativos. A discriminação é, essencialmente, o grau de associação de um item com os restantes e, presumivelmente, com a classificação verdadeira. Habitualmente é expressa como a correlação entre o item em questão e o conjunto de todos os itens, mais propriamente com o conjunto de todos os itens à exceção do próprio item [27].

O termo discriminação deve-se ao facto de um item, que é fortemente correlacionado com o conjunto total de itens, poder discriminar, de forma mais clara, os indivíduos que obtiveram classificações baixas daqueles que obtiveram classificações elevadas no conjunto total de itens [13].

Outra propriedade dos itens é a sua dificuldade. Um item difícil é aquele a que poucos indivíduos conseguem responder corretamente. A dificuldade de um item é quantificada como a proporção de pessoas que respondem corretamente ao item. Obviamente, quanto menor for a proporção de pessoas que respondem corretamente ao item maior é a sua dificuldade [13].

A análise de testes de escolha múltipla com classificações binárias (certo/errado) dos itens usando Teoria Clássica de Testes inclui, portanto, a apresentação da proporção amostral de respostas corretas para cada um dos itens, parâmetro de dificuldade, e dos coeficientes de correlação bisserial entre cada um dos itens e a classificação final excluindo o item em causa, parâmetro de discriminação.

2.2 Teoria de Resposta ao Item (TRI)

A bibliografia correspondente a este capítulo é essencialmente Kolen e Brennan (2004) [28] e Streiner e Norman (2003) [54].

A Teoria de Resposta ao Item é usada em várias áreas onde são aplicados testes e graças sobretudo aos trabalhos desenvolvidos por Wright, Stone, Lord, Hambleton, Swaminathan e Baker e ao desenvolvimento de software informático o seu uso registou um grande progresso nas últimas décadas do século XX [25]. Fazem parte das suas aplicações, entre outras, a criação e desenvolvimento de testes, de bancos de questões, de testes adaptativos e a calibração de testes. Muitas das potencialidades da Teoria de Resposta ao Item resultam do facto de esta modelar a resposta de um indivíduo a cada item ao passo que o foco da Teoria Clássica de Testes é o teste na sua globalidade [28].

Dada a contribuição que a área educacional teve no desenvolvimento dos modelos da Teoria de Resposta ao Item estes são também designados por modelos de traço latente, uma vez que na área educacional as variáveis latentes são concebidas como traços [4]. Os modelos da Teoria de Resposta ao Item consideram os itens cotados dicotomicamente (0/1) ou politomicamente (por exemplo, 0, 1, 2) e podem ser aplicados em testes que pretendem medir uma variável unidimensional ou multidimensional [28]. Estes modelos pressupõem condições estatísticas fortes que nem sempre se verificam na prática, pelo que se revela fundamental analisar as violações dos pressupostos e o ajuste dos modelos aos dados [24], [25].

Esta teoria recorre a modelos não lineares de efeitos mistos que relacionam as respostas observadas com uma variável latente, não observada, atribuindo a cada indivíduo uma estimativa do valor dessa variável latente de acordo com as suas respostas ao teste [44] e determinando o parâmetro de dificuldade dos itens na mesma escala que a da variável latente. Esta escala goza da propriedade de invariância, ou seja, a habilidade de um indivíduo é estimada independentemente do conjunto de itens a que foi submetido e o parâmetro de dificuldade dos itens é estimado independentemente do conjunto de indivíduos que responderam aos itens [15]. No contexto educacional de uma avaliação, a variável latente é usualmente designada por habilidade. A aplicação destes métodos requer a satisfação de duas hipóteses:

- as respostas de um aluno aos itens resultam unicamente do seu nível de habilidade;
- os itens são independentes quando condicionados pela habilidade (condição de independência local).

Em particular, a resposta de um aluno a um item não pode depender da sua resposta a qualquer outro item do teste, uma vez conhecida a habilidade do aluno. A Teoria de Resposta ao Item (TRI) foi usada para avaliar a qualidade dos itens dos testes.

Os modelos de traço latente pressupõem a existência de independência local, dada a habilidade, ou seja, dada a habilidade do indivíduo as respostas aos itens são independentes umas das outras. Na prática esta condição nem sempre se verifica. No caso particular em estudo, a presença nos testes diagnósticos de Matemática da FCUP de itens que abordam o mesmo conteúdo matemático poderá colocar a existência de independência local em causa. Nas situações em que a independência local, dada a habilidade, não se verifica esta condição pode ser relaxada recorrendo, por exemplo, a modelos log-lineares de traço latente com efeitos diretos que têm em consideração as relações entre as respostas aos itens [23].

O modelo da Teoria de Resposta ao Item assume que a habilidade é uma variável contínua seguindo uma distribuição normal reduzida e que a distribuição de cada item i condicionada pela habilidade é uma distribuição de Bernoulli $B(1, \pi_i(\theta))$, $i = 1, 2, \dots, m$ onde $\pi_i(\theta) =$

$P(X_i = 1|\theta)$ é a probabilidade de acertar o item i conhecida a habilidade θ . O gráfico da função π_i é designado por curva característica do item i .

Assumindo que cada item x_i é uma realização de uma variável contínua subjacente x_i^* , a carga factorial a_{i1}^* é o coeficiente de correlação entre a habilidade θ e a variável x_i^* . Neste contexto, é costume classificar-se uma associação como fraca se a sua carga factorial for inferior a 0,3, moderada se se encontra entre 0,3 e 0,7 e forte se for superior a 0,7 [38].

Seja X_{ij} a variável aleatória que representa a resposta do indivíduo j ao item i . Considerando um teste constituído por itens cotados dicotomicamente, tem-se $X_{ij} = 1$ para uma resposta correta ao item e $X_{ij} = 0$ para uma resposta incorreta ao item.

Para o modelo logístico com três parâmetros (3PL) as probabilidades de resposta correta são modeladas da seguinte forma:

$$P(X_{ij} = 1|\theta_j, a_i, b_i, c_i) = c_i + (1 - c_i) \frac{\exp(Da_i(\theta_j - b_i))}{1 + \exp(Da_i(\theta_j - b_i))} \quad (11)$$

onde θ_j é a habilidade do indivíduo j , a constante a_i é designada por parâmetro de discriminação do item i , a constante b_i é designada por parâmetro de dificuldade do item i , a constante c_i é designada por parâmetro de adivinhação (assíntota inferior das curvas características dos itens) que está relacionado com a possibilidade de um estudante indicar a resposta correta a um item por acaso e D é uma constante de escala.

Quando considerado o parâmetro de adivinhação as curvas características dos itens possuem maior valor de interseção com o eixo vertical, designado usualmente por ordenada na origem. Se o parâmetro de adivinhação for nulo a equação (11) corresponde ao modelo logístico com dois parâmetros (2PL) e se, para além disso, os parâmetros de discriminação se considerarem iguais para todos os itens a equação (11) corresponde ao modelo logístico com um parâmetro (1PL), também designado por modelo de Rasch [43], [54], [57].

Os modelos 1PL e 2PL podem ser escritos como modelos logísticos. Para o modelo 1PL, a probabilidade do indivíduo j responder corretamente ao item i é dada por:

$$P_i(\theta_j) = \frac{\exp(a(\theta_j - b_i))}{1 + \exp(a(\theta_j - b_i))} \quad (12)$$

sendo a o parâmetro de discriminação (igual para todos os itens), b_i o parâmetro de dificuldade do item i e θ_j a habilidade do indivíduo j .

Para o modelo 2PL, a probabilidade do indivíduo j responder corretamente ao item i é dada por:

$$P_i(\theta_j) = \frac{\exp(a_i(\theta_j - b_i))}{1 + \exp(a_i(\theta_j - b_i))} \quad (13)$$

sendo a_i o parâmetro de discriminação do item i , b_i o parâmetro de dificuldade do item i e θ_j a habilidade do indivíduo j .

O parâmetro de dificuldade corresponde ao ponto na escala de habilidade para o qual há 50% de probabilidade de acertar o item. Ou seja, para todos os indivíduos cuja habilidade é igual ao parâmetro de dificuldade do item a probabilidade de responder corretamente ao item é de 50%. Se a habilidade dos indivíduos for inferior (resp. superior) ao parâmetro de dificuldade a probabilidade de acertar o item é inferior (resp. superior) a 50% [54].

Partindo do modelo logístico 2PL tem-se:

$$\begin{aligned}
\frac{\exp(a_i(\theta_j - b_i))}{1 + \exp(a_i(\theta_j - b_i))} &= 0,5 \\
\Leftrightarrow \exp(a_i(\theta_j - b_i)) &= 0,5 + 0,5 \exp(a_i(\theta_j - b_i)) \\
\Leftrightarrow \exp(a_i(\theta_j - b_i)) &= 1 \\
\Leftrightarrow a_i(\theta_j - b_i) &= 0 \\
\Leftrightarrow a_i = 0 \vee \theta_j &= b_i
\end{aligned}$$

O parâmetro de discriminação corresponde ao declive da curva característica do item no parâmetro de dificuldade respetivo, indicando portanto a capacidade de discriminação do item para indivíduos com habilidade próxima do parâmetro de dificuldade.

Partindo do modelo logístico 2PL tem-se:

$$P'_i(\theta_j) = \frac{a_i \exp(a_i(\theta_j - b_i))}{(1 + \exp(a_i(\theta_j - b_i)))^2}.$$

Tomando a habilidade do indivíduo igual ao parâmetro de dificuldade do item tem-se:

$$\begin{aligned}
P'_i(b_i) &= \frac{a_i \exp(a_i(b_i - b_i))}{(1 + \exp(a_i(b_i - b_i)))^2} \\
&= \frac{a_i}{4}
\end{aligned}$$

Na realidade o parâmetro de discriminação é proporcional ao declive da curva característica do item no parâmetro de dificuldade respetivo. Considerar que é igual ao declive da curva característica do item no parâmetro de dificuldade respetivo, para além de ser uma boa aproximação, torna a interpretação do parâmetro de discriminação mais fácil na prática [3].

A carga fatorial de um item resulta de padronizar o parâmetro de discriminação do item de forma a obter uma correlação entre o item e o conhecimento subjacente da seguinte forma [7]:

$$a_i^* = \frac{a_i}{\sqrt{a_i^2 + 1}}$$

sendo a_i o parâmetro de discriminação do item i .

É de esperar que as curvas características dos itens sejam funções monótonas crescentes, ou seja, quanto maior a habilidade do indivíduo maior é a probabilidade de este responder corretamente ao item. As curvas características dos itens obtidas com o modelo 1PL têm como característica o facto de não se intersectarem, uma vez que todas as curvas possuem o mesmo parâmetro de discriminação (Figura 1).

Para determinar a quantidade de informação associada a cada curva característica do item recorre-se à curva de informação do item. Segundo Fisher a informação corresponde ao inverso da precisão com que um parâmetro é estimado, ou seja, a informação é o inverso da variância de medição. Quanto maior a precisão com que um parâmetro é estimado maior é a informação sobre o valor do parâmetro [30].

A curva de informação total do teste corresponde à soma de todas as curvas de informação de cada item presente no teste. Esta curva indica a variação da precisão do teste (inverso da variância de medição) em função do valor da habilidade (Figura 2). A partir da curva

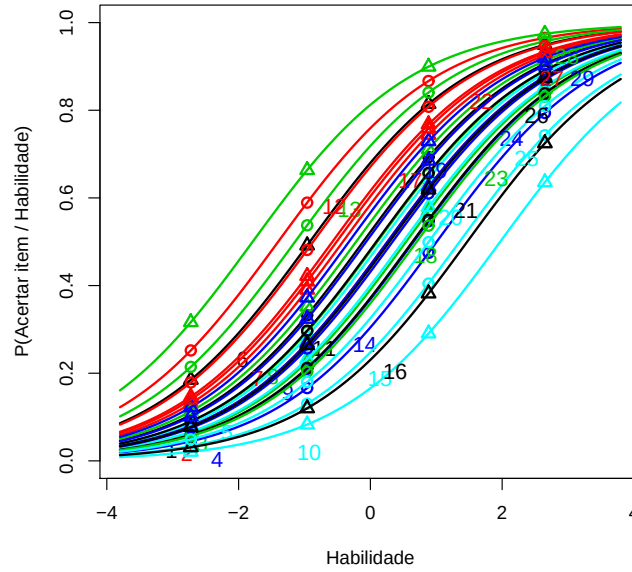


Figura 1: Curvas características dos itens (Modelo 1PL)

de informação, podem também retirar-se intervalos de confiança para a habilidade de cada indivíduo, com erro de medição dependente do padrão de resposta do indivíduo (Figura 3).

Uma vez que o erro padrão de medição não é constante, depende da habilidade do indivíduo, a amplitude dos intervalos de confiança para as classificações finais também varia dependendo da habilidade.

Dada a dimensão das amostras as frequências esperadas para os padrões de resposta são baixas ou nulas. Deparamo-nos, portanto, com um problema de como avaliar a qualidade do ajustamento em tabelas de contingência esparsas [48]. O caso específico em análise foi abordado por Bartholomew e Tzamourani [5]. Estes concluíram que, nesta situação, a melhor forma de proceder a um teste da qualidade do ajustamento global do modelo é gerar empiricamente a distribuição amostral da estatística χ^2 usando o método paramétrico de bootstrap, da seguinte forma:

- 1) Ajustar o modelo aos dados originais e determinar o qui-quadrado (χ^2_{orig});
- 2) Com base nas estimativas dos parâmetros do modelo identificar as distribuições de cada item que, condicionadas pela habilidade θ , seguem uma Bernoulli $B(1, \pi_i(\theta))$, $i=1, 2, \dots, m$;
- 3) Gerar aleatoriamente de $(X_1, \dots, X_m)$ n vetores de resposta, sendo n a dimensão dos dados originais;
- 4) Ajustar o modelo à nova amostra e determinar o qui-quadrado (χ^2_{amost});
- 5) Repetir 3) e 4) um número elevado de vezes (no caso, 500 vezes);
- 6) Considerar o histograma dos vários χ^2_{amost} .

Realizar um teste de hipóteses sendo:

H_0 : A qualidade do ajustamento é boa;

H_1 : A qualidade do ajustamento não é boa.

Para comparar a qualidade do ajustamento dos modelos aos dados seguiu-se um procedi-

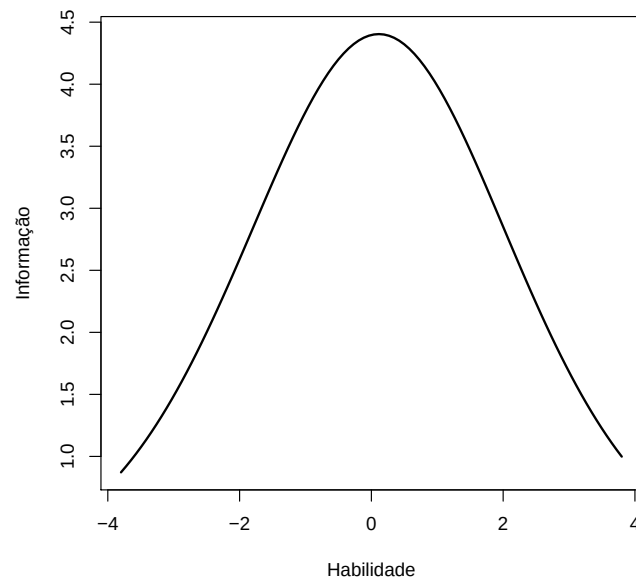


Figura 2: Curva de informação total

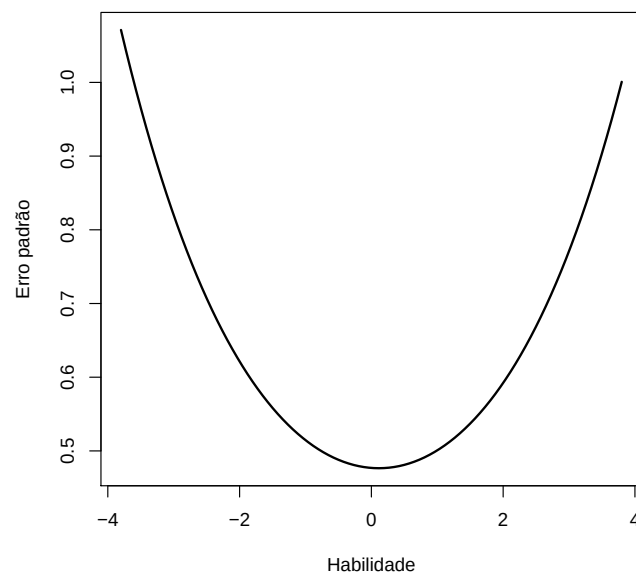


Figura 3: Erro padrão de medição

mento baseado no teste de razão de verossimilhanças (LRT). Este teste é usado para comparar modelos encaixados tendo-se $LRT = -2\ln\left(\frac{L_1}{L_2}\right)$, sendo L_1 a verossimilhança do modelo com menos parâmetros e L_2 a verossimilhança do modelo com mais parâmetros. Assintoticamente a razão de verossimilhanças segue uma distribuição χ^2 com graus de liberdade iguais à diferença entre o número de parâmetros dos modelos [61]. O procedimento seguido foi semelhante ao referido anteriormente para análise da qualidade do ajustamento dos modelos tendo sido realizado um teste de hipóteses considerando:

H_0 : A qualidade do ajustamento dos modelos aos dados é semelhante;

H_1 : O modelo mais complexo revela melhor ajustamento aos dados.

A reduzida dimensão das amostras não permite que seja feito o ajustamento aos dados do modelo logístico com três parâmetros (3PL). Será analisada a qualidade do ajustamento aos dados dos modelos logísticos 1PL e 2PL.

A unidimensionalidade da habilidade foi comprovada através de uma análise em componentes principais sobre a matriz das correlações tetracóricas dos vários itens. A correlação tetracórica é calculada para duas variáveis binárias com variáveis contínuas subjacentes seguindo uma distribuição binormal. Consiste, fundamentalmente, em considerar uma tabela de contingência 2×2 como uma dupla dicotomização de uma distribuição normal bivariada e, em seguida, encontrar o coeficiente de correlação tetracórica de acordo com as probabilidades conjuntas da referida tabela. O coeficiente de correlação tetracórica é, portanto, o coeficiente de correlação linear da designada distribuição normal bivariada latente [16].

A hipótese de independência local será verificada recorrendo à análise dos resíduos marginais de ordem 2.

Ainda no que concerne à análise da qualidade do ajustamento dos modelos recorreu-se à estatística *item-fit* e à estatística *person-fit*. Através da estatística *item-fit* são detetados itens que não se ajustam bem ao modelo em análise. Para tal recorre-se a um teste χ^2 , os indivíduos são ordenados de acordo com a sua habilidade estimada pelo modelo e agrupados de acordo com determinado número de categorias. Dentro de cada grupo é calculada a proporção de indivíduos que respondem corretamente ao item, a qual é comparada com a proporção estimada do modelo da Teoria de Resposta ao Item [51]. De modo semelhante, através da estatística *person-fit* são detetados padrões de resposta aos itens considerados improváveis pelo modelo em análise [40].

A estimação dos parâmetros nos modelos da TRI pode ser feita por máxima verossimilhança ou pelo método dos mínimos quadrados generalizados. Este último método procura estimativas para os parâmetros que minimizem a distância, no sentido dos mínimos quadrados, entre as proporções marginais de primeira e segunda-ordem observadas e esperadas. A estimação por máxima verossimilhança recorre ao algoritmo EM (*Expectation-Maximization*). Usando o teorema de Bayes e algumas manipulações algébricas, os cálculos envolvem apenas a avaliação das probabilidades à posteriori dos padrões de resposta que ocorrem na amostra [4].

2.3 Comparação entre Teoria Clássica de Testes e Teoria de Resposta ao Item

A Teoria de Resposta ao Item apresenta algumas vantagens face à Teoria Clássica de Testes. Uma dessas vantagens é o facto de o erro de medição ser estimado com maior precisão, dependendo da habilidade de cada indivíduo, a precisão das medições é maior para indivíduos com habilidade intermédia do que para indivíduos com habilidades extremas [54]. Na Teoria Clássica de Testes, por seu lado, o erro de medição é comum a todos os indivíduos independentemente da sua habilidade.

Outra das vantagens é o facto das estimativas obtidas serem independentes do teste a que os indivíduos foram submetidos, o que permite a comparação entre grupos de indivíduos submetidos a testes diferentes. Esta propriedade é utilizada para criar bancos de itens, que permitem posicionar os indivíduos, num determinado valor da habilidade, sem que tenham que responder a um número excessivo de itens. Para além da economia de tempo daí resultante, evita-se que indivíduos com baixa (resp. elevada) habilidade se sintam frustrados por terem que responder a um grande número de itens com elevada (resp. baixa) dificuldade. No caso da Teoria Clássica de Testes, um teste com um reduzido número de itens teria associado, provavelmente, um erro padrão de medição elevado. Tal não sucede na Teoria de Resposta ao Item, uma vez que nesta o erro padrão de medição depende apenas da probabilidade de um indivíduo responder corretamente ao item, dada a sua habilidade, pelo que testes com poucos itens podem ser tão ou mais fiáveis do que testes com um maior número de itens [18].

Apesar de serem reconhecidas algumas vantagens da Teoria de Resposta ao Item relativamente à Teoria Clássica de Testes, esta última continua a ser aplicada em várias áreas. Uma das razões que pode justificar este facto resulta de os resultados obtidos com a Teoria Clássica de Testes serem muito próximos dos obtidos com a Teoria de Resposta ao Item quando as amostras têm grande dimensão [54]. Para além disso, a Teoria Clássica de Testes é de mais fácil compreensão pela generalidade das pessoas, não envolvendo conhecimentos matemáticos complexos.

Outra justificação para que a Teoria Clássica de Testes continue a ter vasta aplicação deve-se à propriedade de invariância da Teoria de Resposta ao Item, ou seja, as características dos itens são independentes do grupo de indivíduos que realizaram o teste, ter sido demonstrada teoricamente [24], mas nem sempre se verificar na prática. Em alguns estudos registaram-se diferenças, relativamente significativas, entre as características dos itens considerando amostras de indivíduos diferentes. Tal sugere que a propriedade de invariância não se verifica [9], [41]. É de referir que existem também estudos que comprovaram na prática a propriedade da invariância [25].

Apesar de nos últimos anos terem sido desenvolvidos vários softwares informáticos associados à Teoria de Resposta ao Item, esta teoria apresenta outro constrangimento que tem impedido a sua maior aplicabilidade, a saber, o facto de o pressuposto de unidimensionalidade da habilidade ser bastante forte e, em diversas situações, não se verificar [54].

2.4 Calibração

Comparar as classificações obtidas entre testes revela-se, muitas vezes, uma área de interesse, contudo devido às propriedades psicométricas dos testes utilizados (parâmetro de dificuldade, parâmetro de discriminação, ...) e/ou aos grupos de examinandos e/ou ao enunciado dos testes usados, nem sempre é possível efetuar uma comparação direta dessas classificações, como é costume fazer-se.

Quando se aplicam dois testes, a avaliar o mesmo construto, a dois grupos independentes de indivíduos um dos grupos pode obter melhores classificações do que o outro. Tal pode dever-se ao facto dos indivíduos com melhores classificações serem mais hábeis, ou ao facto de o teste a que foram submetidos ser mais acessível. Para perceber quais as razões que justificam as classificações obtidas é necessário recorrer a técnicas de calibração.

Quando se pretende calibrar classificações observadas em várias administrações de testes o processo pode envolver, de um modo geral, dados de um único grupo de indivíduos que se submeteram aos testes, de grupos equivalentes ou de grupos não equivalentes.

No caso de se ter um único grupo de indivíduos este é submetido a dois testes diferentes, qualquer diferença na distribuição das classificações observadas nos testes é atribuída aos testes em si, em particular, ao grau de dificuldade dos itens que compõem os testes, uma vez que se assume que a habilidade dos indivíduos é constante. Situação semelhante ocorre quando estamos perante grupos equivalentes, ou seja, quando estes foram seleccionados aleatoriamente da população sendo um grupo submetido a um teste e o outro grupo submetido a outro teste. As diferenças nas distribuições das classificações observadas são atribuídas aos testes, uma vez que se assume que os grupos possuem a mesma habilidade. Se os grupos não forem equivalentes a população deve ser definida indiretamente, usando amostras dos dois grupos de indivíduos que se submeteram aos testes, e devem ser tidas em consideração, para além dos testes em si, as diferenças existentes entre a habilidade de cada um dos grupos de indivíduos. Para tal recorre-se a itens âncora presentes em ambos os testes que constituem uma medição da habilidade comum aos dois grupos.

Dado que se considerou que a habilidade dos estudantes, que se submeteram aos testes diagnósticos nos diferentes anos, pode não ser constante estamos perante grupos não equivalentes. Para criar uma escala comum, que permita a comparação das classificações obtidas, é necessário calibrar as diferentes classificações em análise para a escala de referência definida, ou seja, transformar os resultados da escala em que se encontram para uma escala comum a todas as classificações. Este processo é geralmente designado por calibração.

As seguintes condições são necessárias para que se possa realizar o processo de calibração:

- os testes em causa avaliam a mesma matéria (construto);
- calibrar do teste 1 para o teste 2 é igual a calibrar do teste 2 para o teste 1.

O processo de calibração foi realizado seguindo duas metodologias: calibração usando Teoria Clássica de Testes e calibração usando Teoria de Resposta ao Item.

2.4.1 Calibração usando Teoria Clássica de Testes

A Teoria de Resposta ao Item tem sido amplamente utilizada na calibração de resultados de testes, dado que permite grande flexibilidade na forma como a calibração é efetuada,

permitindo calibrar testes com formas bastante díspares. Em determinadas situações é possível recorrer a métodos de calibração mais simples e com pressupostos menos exigentes que se baseiam na Teoria Clássica de Testes [32]. A nível numérico a calibração baseada na Teoria Clássica de Testes é muito menos exigente do que a baseada na Teoria de Resposta ao Item, uma vez que apenas é necessária informação sobre as classificações obtidas pelos indivíduos, ao passo que na calibração usando Teoria de Resposta ao Item é necessário estimar os parâmetros de cada um dos itens.

A leitura de Livingston [32], serviu de base para a explanação do processo de calibração que se segue.

Segundo Livingston [32], uma classificação num novo teste e uma classificação num teste de referência estão calibradas se o mesmo grupo de indivíduos obtiver, no novo teste, a mesma posição relativa que tinha no teste de referência.

No caso em estudo estamos perante grupos não equivalentes, uma vez que se assume que a habilidade dos estudantes que se submeteram aos testes diagnósticos, nos diferentes anos, pode não ser constante.

Aplicámos a calibração separada, ou seja, os parâmetros dos itens dos vários testes foram estimados ano a ano, repetindo-se o processo de calibração para cada um dos anos.

Para realizar a calibração podem ser seguidas várias técnicas, de acordo com a equação que permite transformar as classificações numa escala para a escala de referência, estas designam-se por lineares ou não lineares.

Neste trabalho aplicámos o método linear. Se se considerarem dois testes X e Y diz-se que as classificações observadas x e y, respetivamente no teste X e no teste Y, estão calibradas de forma linear se:

$$\frac{y - \mu_Y}{\sigma_Y} = \frac{x - \mu_X}{\sigma_X} \quad (14)$$

sendo as médias e os desvios padrão calculados no respetivo grupo de indivíduos que se submeteram ao teste [1].

Como estamos perante grupos não equivalentes é necessário que a informação dos itens âncora seja tida em consideração na estimação dos parâmetros anteriormente referidos. Como existem duas populações distintas, cada uma realizando um determinado teste, é necessário definir uma nova população S, que se designa por população sintética [6], à custa das duas populações que se submeteram aos testes.

Considerando a população sintética, a equação (14) pode ser escrita da seguinte forma [28]:

$$Y_S(x) = \frac{\hat{\sigma}_S(Y)}{\hat{\sigma}_S(X)}x - \frac{\hat{\sigma}_S(Y)}{\hat{\sigma}_S(X)}\hat{\mu}_S(X) + \hat{\mu}_S(Y) \quad (15)$$

tendo-se para as médias

$$\begin{aligned} \hat{\mu}_S(X) &= \hat{\mu}_1(X) - w_2\gamma_1[\hat{\mu}_1(V) - \hat{\mu}_2(V)], \\ \hat{\mu}_S(Y) &= \hat{\mu}_2(Y) + w_1\gamma_2[\hat{\mu}_1(V) - \hat{\mu}_2(V)] \end{aligned} \quad (16)$$

e para as variâncias

$$\begin{aligned} \hat{\sigma}_S^2(X) &= \hat{\sigma}_1^2(X) - w_2\gamma_1^2[\hat{\sigma}_1^2(V) - \hat{\sigma}_2^2(V)] + w_1w_2\gamma_1^2[\hat{\mu}_1(V) - \hat{\mu}_2(V)]^2, \\ \hat{\sigma}_S^2(Y) &= \hat{\sigma}_2^2(Y) + w_1\gamma_2^2[\hat{\sigma}_1^2(V) - \hat{\sigma}_2^2(V)] + w_1w_2\gamma_2^2[\hat{\mu}_1(V) - \hat{\mu}_2(V)]^2. \end{aligned} \quad (17)$$

Nas equações anteriores w_1 e w_2 são pesos cuja soma é igual a 1 que especificam a contribuição de cada uma das populações 1 e 2 na estimação. O termo γ representa a relação entre a classificação total em X e em Y com a respetiva classificação no conjunto dos itens âncora, designado por V.

A estimação dos termos γ foi feita recorrendo ao método de Tucker. Neste método a relação entre a classificação total e a obtida nos itens âncora é definida em termos de declives de regressão, γ_1 é o declive da regressão de X em V para a população 1 e γ_2 é o declive da regressão de Y em V para a população 2. Ou seja,

$$\begin{aligned}\gamma_1 &= \frac{\hat{\sigma}_1(X, V)}{\hat{\sigma}_1^2(V)} \\ \gamma_2 &= \frac{\hat{\sigma}_2(Y, V)}{\hat{\sigma}_2^2(V)}.\end{aligned}\tag{18}$$

A calibração linear pode levar a que classificações muito elevadas ou muito baixas resultem em valores para as classificações calibradas para a escala de referência fora dessa escala. Outro problema da calibração linear resulta de esta depender fortemente do grupo de indivíduos que se submeteu ao teste cujas classificações se pretendem calibrar. Quando os dois testes diferem de forma significativa no seu grau de dificuldade a calibração linear pode diferir consideravelmente quando considerado um grupo de indivíduos com habilidade elevada e um com reduzida habilidade [32].

Para fazer face a esta situação pode recorrer-se a técnicas de calibração não linear, em particular, à calibração por igualdade de percentis. Na calibração por igualdade de percentis é definida uma relação não linear entre as duas escalas, de modo a que exista uma igualdade entre as distribuições cumulativas das classificações em cada uma das escalas [1].

Neste estudo não usámos calibração não linear devido à reduzida dimensão das amostras e ao diminuto número de itens âncora entre os testes de 2009 e 2010 e os dos restantes anos.

2.4.2 Calibração usando Teoria de Resposta ao Item

Os modelos assentes na Teoria de Resposta ao Item verificam a propriedade de invariância dos parâmetros dos itens, ou seja, as estimativas dos parâmetros dos itens são independentes da habilidade dos indivíduos que a eles se submeteram [17]. Para a realização do processo de calibração recorreu-se a esta propriedade.

Quando se pretende calibrar as classificações de mais do que dois testes é necessário determinar um conjunto de constantes de ligação para cada par de testes adjacentes. Estas constantes transformam os parâmetros de um teste da escala em que se encontram para a escala do outro teste sendo, posteriormente, calibrados para uma escala comum a todos os testes. Tal como na Teoria Clássica de Testes aplicámos a calibração separada.

A leitura de Weeks [60], serviu de base para a explanação do processo de calibração que se segue.

Tendo em consideração os modelos logísticos em análise e o facto das questões serem cotadas dicotomicamente tem-se que, para um determinado par de testes, a equação que permite transformar θ_O (escala original) em θ_F (escala final) é:

$$\theta_{F,i} = A\theta_{O,i} + B = \theta_{O,i}^*\tag{19}$$

para o indivíduo i . Tal como atrás, usaremos $\theta_F = (\theta_{F,1}, \dots, \theta_{F,n})$ e $\theta_O = (\theta_{O,1}, \dots, \theta_{O,n})$ para os vetores das classificações finais e originais, respetivamente. A e B são as constantes usadas para ajustar o desvio padrão e a média, respetivamente, e $*$ representa o valor obtido na escala final.

A alteração da escala leva a uma alteração dos parâmetros dos itens de modo a que as probabilidades esperadas se mantenham inalteradas, pelo que se verifica:

$$\begin{aligned} a_{jO}^* &= \frac{a_{jO}}{A} \\ b_{jO}^* &= Ab_{jO} + B \end{aligned} \quad (20)$$

sendo j o item e A e B as constantes presentes na equação (19). Se se considerar o modelo logístico com três parâmetros (3PL), o parâmetro de adivinhação não é afetado pela transformação pelo que $c_{jO}^* = c_{jO}$.

No caso de se pretender ter em consideração uma calibração as equações (19) e (20) podem ser reescritas de modo a transformar os parâmetros da escala final para a escala original obtendo-se:

$$\begin{aligned} a_{jF}^\# &= Aa_{jF} \\ b_{jF}^\# &= \frac{b_{jF} - B}{A} \end{aligned} \quad (21)$$

denotando $\#$ o valor obtido na escala original. Tem-se ainda $c_{jF}^\# = c_{jF}$, uma vez que o parâmetro de adivinhação não sofre qualquer alteração.

Considerando J o conjunto de todos os itens presentes nos vários testes o subconjunto V de J dos itens que são comuns a todos os testes, designados itens âncora, é usado para estimar as constantes A e B .

Existem vários métodos de calibração baseados na Teoria de Resposta ao Item [28], [60]. Os quatro métodos mais usados são média/desvio padrão [37], média/média [34], Haebara [22] e Stocking-Lord [52]. Os métodos média/desvio padrão e média/média, conhecidos por métodos dos momentos, são os que permitem determinar as referidas constantes de forma mais simples, uma vez que apenas implicam o cálculo das médias e desvios padrão dos diferentes parâmetros dos itens.

No método média/desvio padrão as médias e os desvios padrão são calculados considerando apenas os itens âncora tendo-se:

$$\begin{aligned} A &= \frac{\sigma(b_F)}{\sigma(b_O)} \\ B &= \mu(b_F) - A\mu(b_O). \end{aligned} \quad (22)$$

Ao contrário do método anterior que não tem em consideração o parâmetro de discriminação o método média/média, aquele que aplicámos neste estudo, considera os dois parâmetros na estimação das constantes de ligação as quais se obtêm da seguinte forma:

$$\begin{aligned} A &= \frac{\mu(a_O)}{\mu(a_F)} \\ B &= \mu(b_F) - A\mu(b_O). \end{aligned} \quad (23)$$

Haebara (1980) e Stocking e Lord (1983) propuseram métodos baseados nas curvas características para estimação das constantes de ligação, as quais se obtêm minimizando a soma do quadrado das diferenças entre as curvas características dos itens âncora. Ambos os métodos assumem que os parâmetros na escala final são estimados sem erro. Para fazer face a este facto Haebara propôs um novo critério que considera simultaneamente uma transformação dos parâmetros da escala original para a escala final e vice-versa. Esta abordagem é conhecida como simétrica e a abordagem típica como não simétrica.

Capítulo 3

Aplicação das metodologias aos dados dos testes

Neste capítulo, é apresentada uma análise cuidada e o mais completa possível dos resultados obtidos dos testes diagnósticos levados a cabo na FCUP entre os anos 2009 e 2013. A análise será feita primeiro numa base anual e depois numa base longitudinal para comparação das habilidades médias das diferentes coortes de entrada recorrendo, para o efeito, a técnicas de calibração. Para os anos de 2009 e 2013, foram elaborados relatórios, pela Prof Ana Rita Gaio e pelos Prof Christian Lomp e Prof Paula Lomp, respetivamente, que naturalmente funcionaram também como referências bibliográficas para o trabalho agora apresentado.

3.1 Ano 2009

3.1.1 Descrição dos dados/Análise descritiva e exploratória

As perguntas do teste diagnóstico foram organizadas da forma indicada na Tabela 1.

O item 6 foi eliminado no início do teste por conter, por lapso, duas opções iguais como resposta possível. Todas as análises estatísticas que se seguem consideraram apenas os 29 itens restantes.

No teste em causa, que foi realizado por 363 estudantes, a classificação final podia tomar valores entre 0 e 29 pontos. Para simplificação da apresentação das classificações, a nota final foi considerada na escala usual de 0-20 valores.

3.1.2 Resultados

Uma vez que foi decidido pela comissão responsável pela implementação dos testes diagnósticos não proceder a descontos na situação de indicação de resposta errada, a grande maioria dos alunos optou por responder a quase todas as questões que lhes foram colocadas no teste.

Responderam corretamente a mais de metade dos itens 153 estudantes (42%). A distribuição observada para as classificações (0-20 valores) foi razoavelmente simétrica, com uma mediana de 9,0 valores, uma média de 9,6 valores, e quartis de 7,6 e 11,7 valores. A classificação máxima atribuída foi de 20 valores (a um aluno) e a mínima foi de 0 valores

P1	Simplificação de expressões (quadrado do binómio)
P2	Área de figuras planas
P3	Regiões do plano
P4	Simplificação de expressões com módulos
P5	Razões trigonométricas
P6	Equações de retas perpendiculares no plano
P7	Inequações com módulos
P8	Simplificação de expressões irracionais
P9	Simplificação de expressões irracionais
P10	Simplificação de expressões exponenciais e logarítmicas
P11	Inequações irracionais
P12	Equações exponenciais
P13	Propriedades dos logaritmos
P14	Função composta
P15	Derivada de uma função irracional
P16	Derivada de uma função exponencial
P17	Cálculo de limites
P18	Cálculo de limites
P19	Equações do segundo grau
P20	Características de uma função quadrática
P21	Gráfico de uma função racional
P22	Propriedades das funções racionais
P23	Função composta
P24	Domínio de uma função
P25	Gráfico de uma função dada a sua derivada
P26	Interseção dos gráficos de duas funções
P27	Simplificação de expressões trigonométricas
P28	Extremos de uma função trigonométrica
P29	Gráfico da derivada de uma função
P30	Inversa de uma função

Tabela 1: Organização dos itens (2009)

(a 2 alunos). Identificaram-se 11 *outliers*, 4 com classificações demasiado baixas e 7 com classificações demasiado altas (Figura 4).

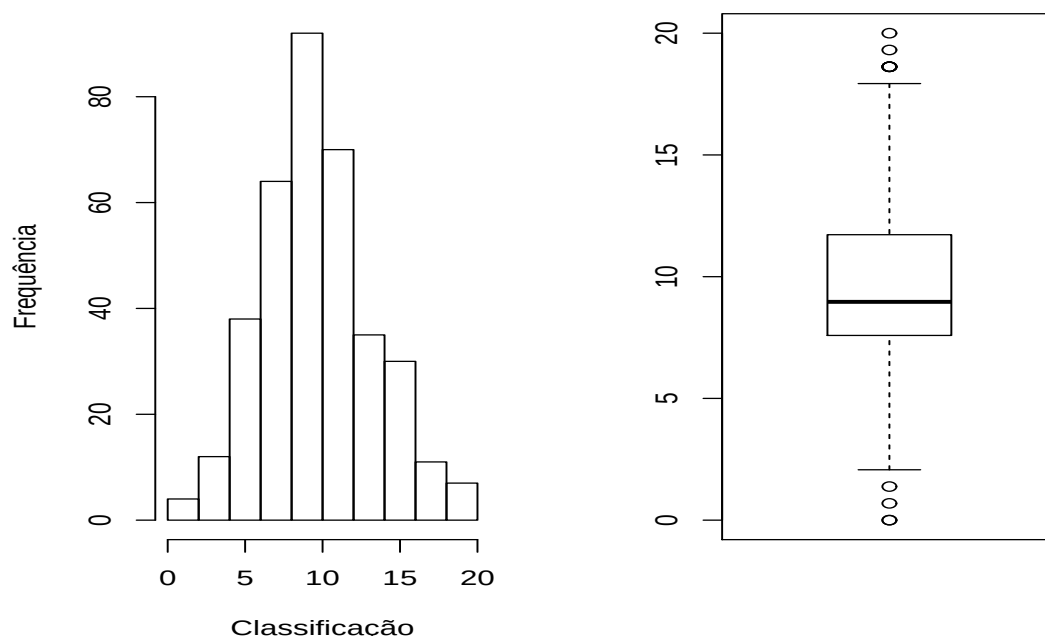


Figura 4: Classificações obtidas (2009)

3.1.2.1 Teoria Clássica de Testes (TCT)

Determinou-se a proporção amostral de respostas corretas para cada um dos itens e os coeficientes de correlação bisserial entre cada um dos itens e a classificação final excluindo o item em causa (Tabela 2). Em relação às percentagens de respostas corretas por item, a mediana foi de 47%, a média foi de 48%, os quartis de 39% e 57%, o mínimo foi de 19%, atingido na questão 11, e o máximo foi de 79%, atingido na questão 29 (Figura 5).

Quanto aos coeficientes de correlação bisserial entre cada um dos itens e a classificação final excluindo o item em causa, o valor mínimo encontrado foi de -0,158, na questão 26, e o máximo foi de 0,532, observado na questão 8. A média destas várias correlações foi de 0,390, a mediana foi de 0,433 e os quartis de 0,353 e 0,466. Com exceção do coeficiente negativo e de um outro igual a apenas 0,102 (item 15), os valores encontrados estão dentro do que é usual encontrar-se na teoria de testes para avaliação educacional [59] (Tabela 2).

A consistência interna do teste foi aferida através do coeficiente alfa de Cronbach, avaliado em 0,791, e do erro padrão de medição, avaliado em 2,43. A amplitude dos intervalos a 95% de confiança para a classificação verdadeira foi de 6,5 valores, numa escala de 0 a 20 valores.

A avaliação da possível existência de itens não consistentes foi feita, para cada item, através do cálculo do coeficiente alfa de Cronbach para o conjunto de itens restantes após o item em causa ter sido eliminado. Apenas a eliminação da pergunta 15 e da pergunta 26 conduziu a um pequeno aumento do valor do coeficiente alfa de Cronbach correspondente à totalidade das perguntas (Tabela 2). Para além deste facto, como o item 26 apresenta

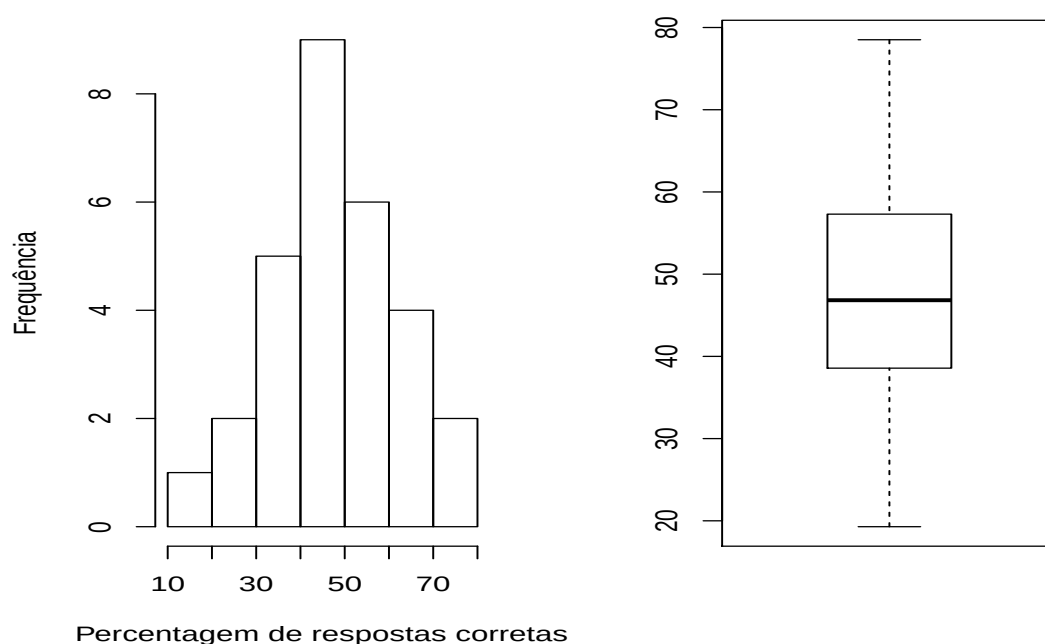


Figura 5: Percentagem de respostas corretas - para os itens (2009)

parâmetro de discriminação negativo e o item 15 é pouco discriminativo, ambos os itens devem ser eliminados.

3.1.2.2 Teoria de Resposta ao Item (TRI)

Para análise da qualidade do ajustamento dos modelos logísticos com um parâmetro (1PL) e com dois parâmetros (2PL) realizou-se o procedimento anteriormente referido no capítulo Métodos (pág. 24). Definindo como nível de significância dos respetivos testes de hipóteses 5% concluiu-se pela não rejeição da hipótese nula, pelo que quer o modelo logístico com um parâmetro (1PL) quer o de dois parâmetros (2PL) revelaram bom ajustamento aos dados. Não foram, portanto, identificadas diferenças estatisticamente significativas entre os dois modelos, no que concerne à qualidade do ajustamento.

No caso do modelo 1PL o percentil 95 da distribuição da estatística do teste do qui-quadrado gerado por *bootstrap* foi de 779161623 e nos dados originais foi de 431006004 (Tabela 3 e Figura 6).

No caso do modelo 2PL o percentil 95 da distribuição da estatística do teste do qui-quadrado gerado por *bootstrap* foi de 781677043 e nos dados originais foi de 419514885 (Tabela 4 e Figura 7).

Para comparação da qualidade do ajustamento dos modelos 1PL e 2PL realizou-se o teste da razão de verossimilhanças, concluindo-se pela não rejeição da hipótese nula, pelo que não existe evidência estatística de que o modelo 2PL se ajuste significativamente melhor aos dados do que o modelo 1PL. Razão pela qual se deveria reter este último modelo (1PL), por ser o mais simples.

Item	TCT			TRI		
	p(%)	Correlação bisserial	α eliminando o item	Parâmetro de dificuldade	Parâmetro de discriminação	Fator de carga
P1	52	0,440	0,783	-0,103	1,027	0,716
P2	45	0,359	0,786	0,299	0,690	0,568
P3	48	0,353	0,786	0,113	0,741	0,595
P4	33	0,438	0,784	0,926	0,906	0,672
P5	47	0,466	0,782	0,137	0,975	0,698
P7	66	0,433	0,784	-0,779	1,034	0,719
P8	57	0,532	0,780	-0,299	1,491	0,831
P9	54	0,349	0,787	-0,244	0,710	0,579
P10	45	0,393	0,785	0,257	0,813	0,631
P11	19	0,466	0,785	1,670	1,018	0,713
P12	48	0,489	0,781	0,054	1,178	0,762
P13	73	0,462	0,784	-1,090	1,162	0,758
P14	69	0,480	0,783	-0,801	1,393	0,812
P15	38	0,102	0,795	2,792	0,172	0,170
P16	27	0,395	0,786	1,408	0,779	0,614
P17	26	0,336	0,788	1,730	0,672	0,558
P18	60	0,438	0,783	-0,510	0,983	0,701
P19	40	0,485	0,782	0,411	1,144	0,753
P20	56	0,386	0,785	-0,315	0,874	0,658
P21	41	0,355	0,786	0,566	0,694	0,570
P22	39	0,429	0,784	0,596	0,884	0,662
P23	65	0,511	0,781	-0,600	1,458	0,825
P24	37	0,435	0,784	0,673	0,860	0,652
P25	44	0,338	0,787	0,385	0,696	0,572
P26	34	-0,158	0,804	-1,943	-0,341	-0,323
P27	45	0,328	0,787	0,361	0,634	0,535
P28	59	0,264	0,790	-0,876	0,434	0,398
P29	79	0,506	0,783	-1,171	1,528	0,837
P30	51	0,499	0,781	-0,064	1,171	0,760
Média	48	0,390	0,785	0,123	0,889	0,621

Tabela 2: Parâmetros estimados usando Teoria Clássica de Testes e Teoria de Resposta ao Item - Modelo 2PL (2009)

p_0	p_{25}	p_{50}	p_{75}	p_{95}	p_{100}
340804470	458981846	515018189	596426523	779161623	1916467359

Tabela 3: Percentis do χ^2 - Modelo 1PL (2009)

p_0	p_{25}	p_{50}	p_{75}	p_{95}	p_{100}
315113138	421587268	485306380	578205517	781677043	1564443486

Tabela 4: Percentis do χ^2 - Modelo 2PL (2009)

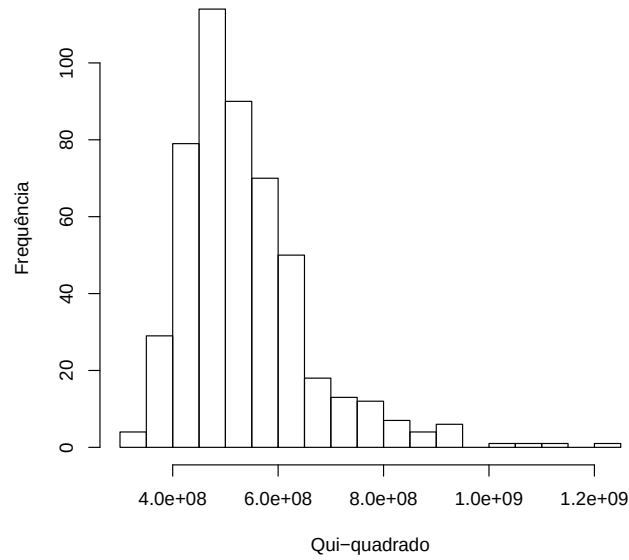


Figura 6: Estatísticas de teste do χ^2 - Modelo 1PL (2009)

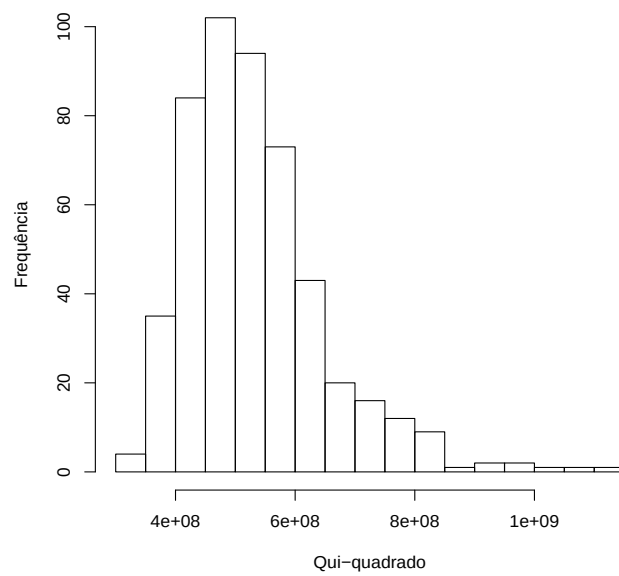


Figura 7: Estatísticas de teste do χ^2 - Modelo 2PL (2009)

O percentil 95 da distribuição da estatística do teste da razão de verossimilhanças gerado por *bootstrap* foi de 260,8 sendo que nos dados originais foi de 172,6 (Tabela 5 e Figura 8).

p_0	p_{25}	p_{50}	p_{75}	p_{95}	p_{100}
117,8	176,4	198,2	221,3	260,8	310,9

Tabela 5: Percentis das razões de verossimilhança (2009)

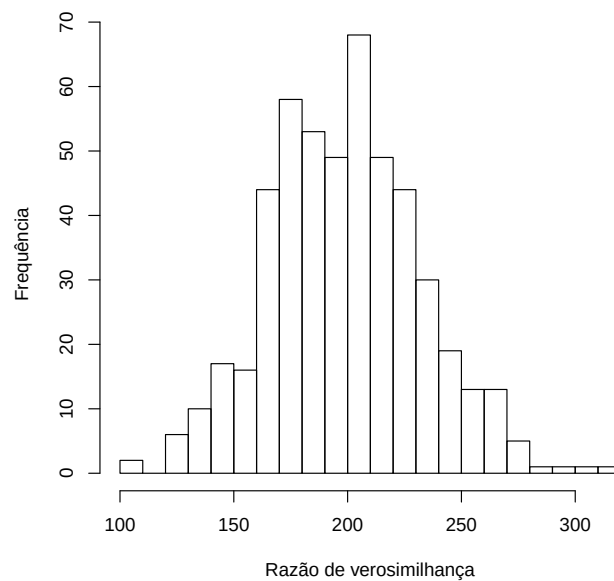


Figura 8: Estatísticas de teste da razão de verossimilhanças (2009)

A qualidade do ajustamento do modelo e a hipótese de independência local foram avaliadas analisando os resíduos marginais de ordem 2.

No caso do modelo 1PL detetaram-se diferenças significativas (resíduos qui-quadrado maiores do que 3,5) entre as frequências observada e esperada em alguns padrões de resposta de pares de itens, a saber: (item 11, item 26), (item 12, item 26), (item 14, item 23), (item 14, item 26), (item 23, item 26), (item 23, item 29) e (item 26, item 29), Tabela 6.

Pares de itens	(0,0)		(1,0)		(0,1)		(1,1)	
	O	E	O	E	O	E	O	E
(11,26)	-	-	53	39	-	-	17	31
(12,26)	-	-	-	-	-	-	51	72
(14,23)	80	48	47	79	31	63	205	174
(14,26)	63	81	-	-	48	29	-	-
(23,26)	-	-	-	-	54	34	-	-
(23,29)	49	35	-	-	-	-	-	-
(26,29)	-	-	34	20	-	-	-	-

Tabela 6: Avaliação marginal da qualidade do ajustamento - Modelo 1PL (2009)

A avaliação marginal da qualidade do ajustamento do modelo revela, portanto, qualidade insatisfatória, não se verificando a hipótese de independência local.

No caso do modelo 2PL apenas se detetaram diferenças significativas entre as frequências observada e esperada dos padrões de resposta para os pares de itens (item 8, item 24) e (item 14, item 23), ambos relacionados com a composição de funções. Verificou-se, por exemplo, que os alunos que erraram simultaneamente as duas questões o fizeram um número de vezes (80) superior ao previsto pelo modelo (59), Tabela 7.

Pares de itens	(0,0)		(1,0)		(0,1)	
	O	E	O	E	O	E
(8,24)	-	-	-	-	56	41
(14,23)	80	59	47	67	31	-

Tabela 7: Avaliação marginal da qualidade do ajustamento - Modelo 2PL (2009)

A avaliação marginal da qualidade do ajustamento do modelo revela, portanto, boa qualidade, assumindo-se como verificada a hipótese de independência local.

Pelo exposto, neste estudo, retivemos o modelo 2PL acima descrito. O histograma das habilidades previstas é apresentado na Figura 9 e as estimativas obtidas para os parâmetros de dificuldade e discriminação são apresentadas na Tabela 2.

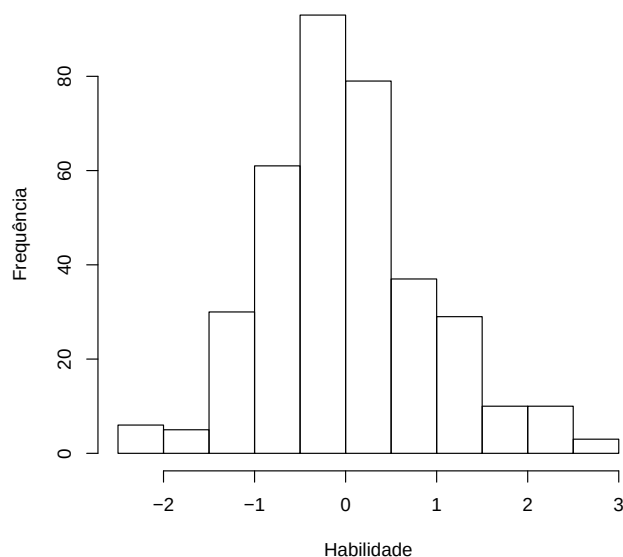


Figura 9: Habilidades previstas pelo modelo da TRI (2009)

Os parâmetros de dificuldade variaram entre -1,943 e 2,792, apresentando uma mediana de 0,113; os parâmetros de discriminação variaram entre -0,341 e 1,528, com uma mediana de 0,884. Repare-se contudo que se encontrou um (único) parâmetro de discriminação negativo, correspondente ao item 26, e que o percentil 25 foi já de 0,697. Num item com parâmetro de discriminação negativo a probabilidade de resposta correta ao item diminui

com o aumento da habilidade do indivíduo. Geralmente, esta situação ocorre quando o item não está bem formulado ou existe alguma informação que induz em erro os indivíduos com maior habilidade.

Também foram incluídas na Tabela 2 as cargas fatoriais estimadas as quais variaram entre -0,323 e 0,837, com uma mediana de 0,663 e percentil 25 de 0,572. Este comportamento é análogo ao dos parâmetros de discriminação, dado que quanto maior for o declive de uma curva característica de um item, maior é a correlação entre esse item e a habilidade, uma vez que quanto maior for o parâmetro de discriminação de um item maior é a capacidade do item de distinguir indivíduos com habilidades próximas do parâmetro de dificuldade. As curvas características dos itens são apresentadas na Figura 10. É possível observar que o item 26 apresenta um parâmetro de discriminação negativo e que os itens 15 e 28 apresentam baixa capacidade de discriminação. A combinação dos elementos apresentados na Tabela 2 e Figura 10 sugere que: o item 26 não deveria constar da escala de avaliação final, por apresentar uma carga factorial negativa; os itens 15 e 28 são pouco discriminativos, e poderiam eventualmente ser também eliminados da escala. As respostas obtidas por categoria de opção para cada um destes itens são apresentadas na Tabela 8. Verifica-se que no item 15 e no item 26 houve, respetivamente, duas e três opções muito disputadas pelos alunos. A questão 15 dizia respeito ao cálculo da primeira derivada de uma função irracional, apresentou elevado grau de dificuldade tendo os alunos hesitado entre duas das opções de resposta (Tabela 2 e Tabela 8). Para além de, como já foi referido, ter apresentado baixa capacidade de discriminação entre os indivíduos. Relativamente à questão 26 dizia respeito ao número de interseções do gráfico de uma dada função trigonométrica com uma dada função exponencial. Nesta questão os alunos com maior habilidade revelaram maior dificuldade em selecionar a resposta correta do que os alunos com menor habilidade. A distribuição das respostas a este item dividindo os indivíduos com maior habilidade dos que têm menor habilidade é semelhante. O maior acerto dos estudantes com menor habilidade deveu-se ao acaso. Na questão 28 pretendia-se que os estudantes indicassem o número de extremos de uma dada função trigonométrica num determinado intervalo.

Item	NR	A	B	C	D
15	20 (6%)	139 (38%)	36 (10%)	152 (42%)	16 (4%)
26	13 (4%)	16 (4%)	125 (34%)	92 (25%)	117 (32%)
28	19 (5%)	28 (8%)	72 (20%)	214 (59%)	30 (8%)

Tabela 8: Respostas aos itens 15, 26 e 28 por opção (A, B, C ou D), para o ano de 2009. As células a verde representam a opção correta.

A curva de informação total do teste é apresentada na Figura 11.

O gráfico do erro padrão de medição em função da habilidade corresponde à Figura 12. Em particular, estes gráficos (equivalentes) sugerem valores da habilidade para os quais a fiabilidade do teste é maior. Repare-se que a zona de maior precisão corresponde precisamente aos valores da habilidade para os quais já se tinha identificado um maior poder de discriminação dos itens. O ponto máximo de informação (5,966) foi atingido em 0,342 desvios padrão abaixo da média da habilidade. Este valor corresponde a um coeficiente alfa de Cronbach superior ao valor (constante) obtido da TCT.

Relativamente à unidimensionalidade da habilidade esta foi comprovada através de uma

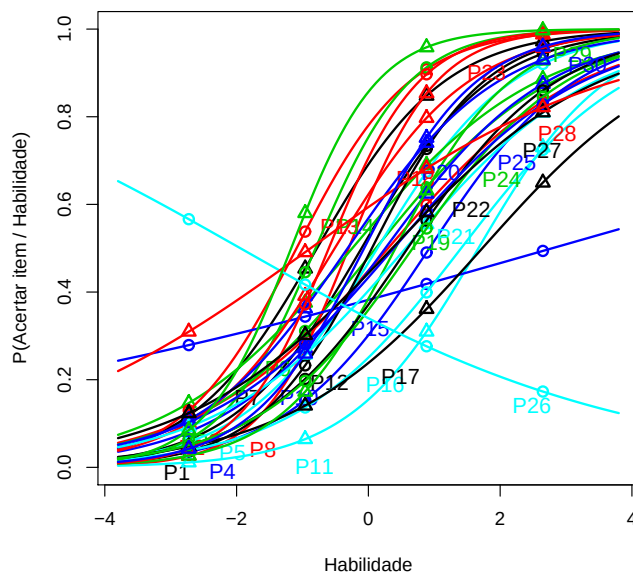


Figura 10: Curvas características dos itens (2009)

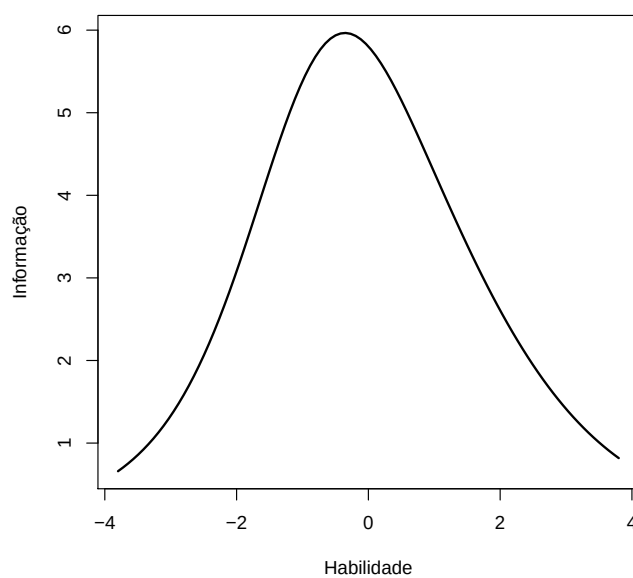


Figura 11: Curva de informação total (2009)

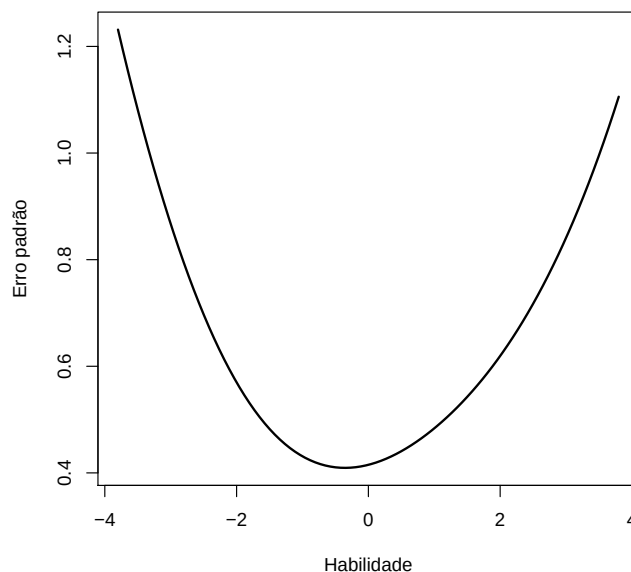


Figura 12: Erro padrão de medição (2009)

análise em componentes principais sobre a matriz das correlações tetracóricas dos vários itens. A percentagem da variância total explicada pela primeira componente foi de 23,3% enquanto que essa percentagem foi de apenas 5,6% para a segunda componente (Figura 13).

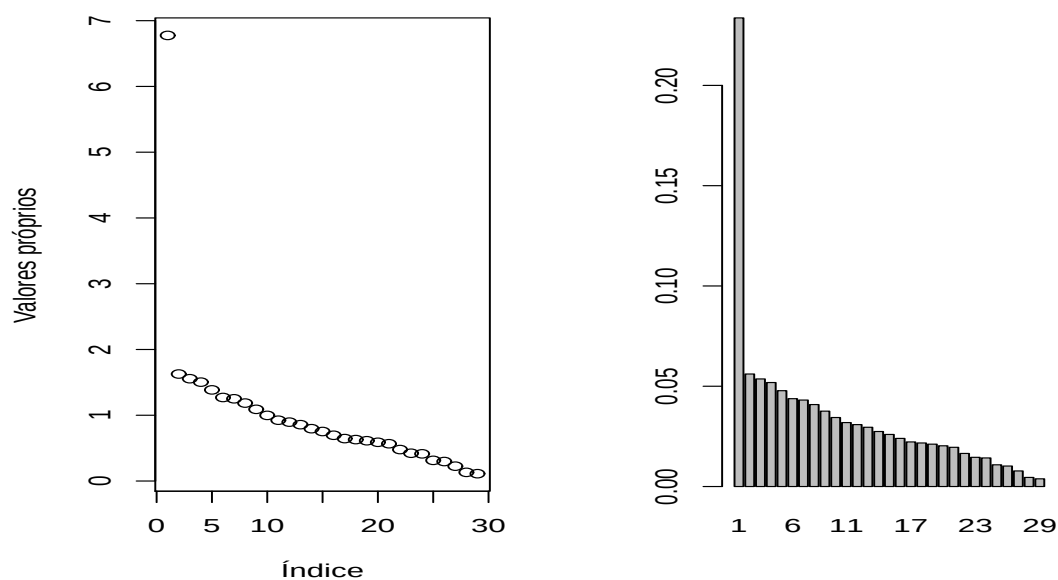


Figura 13: Valores próprios e componentes principais (2009)

A estatística *item-fit* mostrou que apenas 6 itens (14, 16, 22, 23, 27, 29) não se ajustaram

corretamente ao modelo seguido, enquanto que a estatística *person-fit* identificou apenas 7 padrões de resposta como inconsistentes para o modelo estimado.

3.2 Ano 2010

3.2.1 Descrição dos dados/Análise descritiva e exploratória

No ano 2010 foi aplicado o mesmo teste diagnóstico do ano 2009, tendo sido feita a correção das opções da questão 6 e substituída a questão 26, que se revelara problemática, por outra questão (Apêndice 1). Realizaram o teste 218 estudantes. Seguiu-se a mesma abordagem na análise dos resultados obtidos.

3.2.2 Resultados

Responderam corretamente a pelo menos metade dos itens 87 estudantes (40%). A distribuição observada para as classificações (0-20 valores) foi razoavelmente simétrica, com uma mediana de 8,7 valores, uma média de 9,1 valores, e quartis de 6,7 e 11,3 valores. A classificação máxima atribuída foi de 19,3 valores (a dois alunos) e a mínima foi de 0 valores (obtida por um aluno), Figura 14. Identificaram-se sete *outliers*, um com uma classificação demasiado baixa e 6 com classificações demasiado altas.

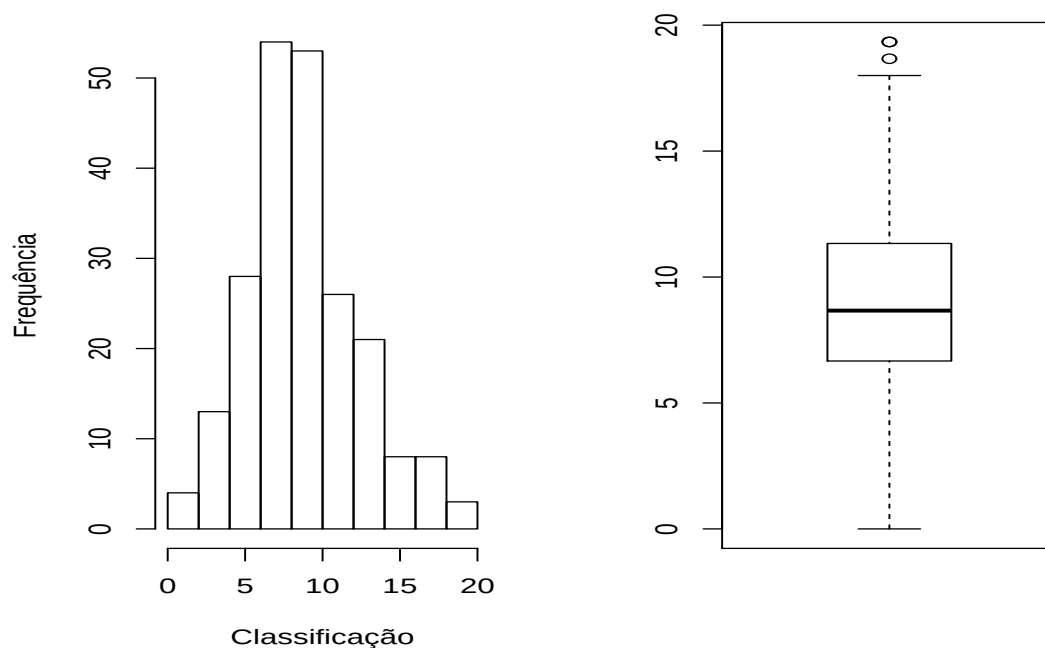


Figura 14: Classificações obtidas (2010)

3.2.2.1 Teoria Clássica de Testes (TCT)

Em relação às percentagens de respostas corretas por item, a mediana foi de 46% assim como a média, os quartis de 37% e 55%, o mínimo foi de 15%, atingido na questão 11, e o máximo foi de 75%, atingido na questão 29 (Figura 15).

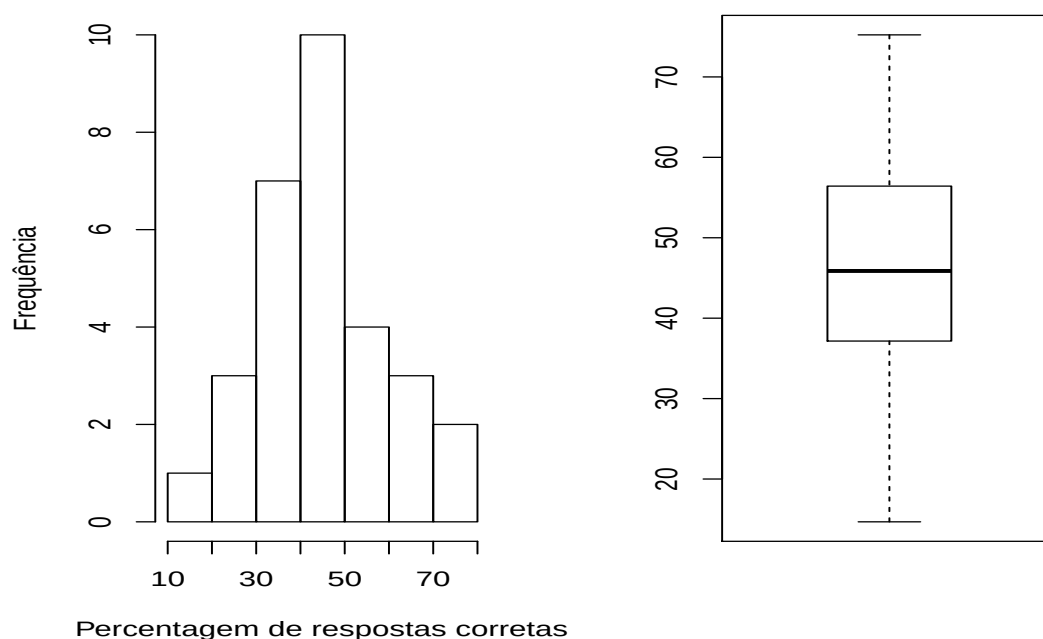


Figura 15: Percentagem de respostas corretas - para os itens (2010)

Quanto aos coeficientes de correlação bisserial entre cada um dos itens e a classificação final excluindo o item em causa, o valor mínimo encontrado foi de 0,107, na questão 15, e o máximo foi de 0,558, observado na questão 26. A média destas várias correlações foi de 0,403, a mediana foi de 0,420 e os quartis de 0,366 e 0,447. Com exceção do coeficiente mínimo e do obtido para o item 25 (0,177) os valores encontrados estão dentro do que é usual encontrar-se na teoria de testes para avaliação educacional (Tabela 9).

A consistência interna do teste foi aferida através do coeficiente alfa de Cronbach, avaliado em 0,802, e do erro padrão de medição, avaliado em 2,47. A amplitude dos intervalos a 95% de confiança para a classificação verdadeira foi de 6,3 valores, numa escala de 0 a 20 valores.

A avaliação da possível existência de itens não consistentes foi feita, para cada item, através do cálculo do coeficiente alfa de Cronbach para o conjunto de itens restantes após o item em causa ter sido eliminado. Apenas a eliminação da pergunta 15 e da pergunta 25 conduziu a um pequeno aumento do valor do coeficiente alfa de Cronbach correspondente à totalidade das perguntas (Tabela 9). Como estas questões são também pouco discriminativas devem ser eliminadas.

Item	TCT			TRI		
	p(%)	Correlação bisserial	α eliminando o item	Parâmetro de dificuldade	Parâmetro de discriminação	Fator de carga
P1	47	0,419	0,796	0,147	0,889	0,664
P2	46	0,361	0,798	0,227	0,766	0,608
P3	43	0,432	0,795	0,331	0,920	0,677
P4	28	0,337	0,799	1,449	0,697	0,572
P5	49	0,402	0,796	0,054	0,864	0,654
P6	44	0,402	0,796	0,311	0,833	0,640
P7	62	0,450	0,795	-0,573	1,081	0,734
P8	57	0,529	0,792	-0,317	1,395	0,813
P9	52	0,457	0,794	-0,132	1,044	0,722
P10	43	0,432	0,795	-0,319	0,960	0,693
P11	15	0,300	0,801	3,288	0,569	0,495
P12	47	0,450	0,795	0,136	0,955	0,691
P13	69	0,407	0,797	-0,980	0,969	0,696
P14	62	0,408	0,796	-0,666	0,864	0,654
P15	35	0,107	0,806	2,670	0,229	0,223
P16	27	0,433	0,796	1,294	0,874	0,658
P17	24	0,282	0,801	2,363	0,519	0,460
P18	56	0,378	0,797	-0,377	0,820	0,634
P19	37	0,420	0,796	0,728	0,801	0,625
P20	47	0,362	0,798	0,149	0,748	0,599
P21	38	0,272	0,801	1,094	0,484	0,436
P22	35	0,427	0,796	0,799	0,887	0,664
P23	56	0,463	0,794	-0,315	1,101	0,740
P24	33	0,446	0,795	0,917	0,875	0,659
P25	39	0,177	0,804	1,414	0,324	0,308
P26	72	0,558	0,792	-0,789	1,804	0,875
P27	39	0,442	0,795	0,561	0,897	0,668
P28	46	0,377	0,797	0,223	0,779	0,614
P29	75	0,503	0,795	-0,982	1,631	0,853
P30	48	0,444	0,795	0,083	1,020	0,714
Média	46	0,403	0,797	0,448	0,887	0,635

Tabela 9: Parâmetros estimados usando Teoria Clássica de Testes e Teoria de Resposta ao Item - Modelo 2PL (2010)

3.2.2.2 Teoria de Resposta ao Item (TRI)

Para análise da qualidade do ajustamento dos modelos logísticos com um parâmetro (1PL) e com dois parâmetros (2PL) realizou-se o procedimento anteriormente referido no capítulo Métodos (pág. 24). Definindo como nível de significância dos respectivos testes de hipóteses 5% concluiu-se pela não rejeição da hipótese nula, pelo que quer o modelo logístico com um parâmetro (1PL) quer o de dois parâmetros (2PL) revelaram bom ajustamento aos

dados. Não foram, portanto, identificadas diferenças estatisticamente significativas entre os dois modelos, no que concerne à qualidade do ajustamento.

No caso do modelo 1PL o percentil 95 da distribuição da estatística do teste do qui-quadrado gerado por *bootstrap* foi de 1803178766 e nos dados originais foi de 1367130274 (Tabela 10 e Figura 16).

p_0	p_{25}	p_{50}	p_{75}	p_{95}	p_{100}
572338966	839134177	991057331	1204401348	1803178766	6274116050

Tabela 10: Percentis do χ^2 - Modelo 1PL (2010)

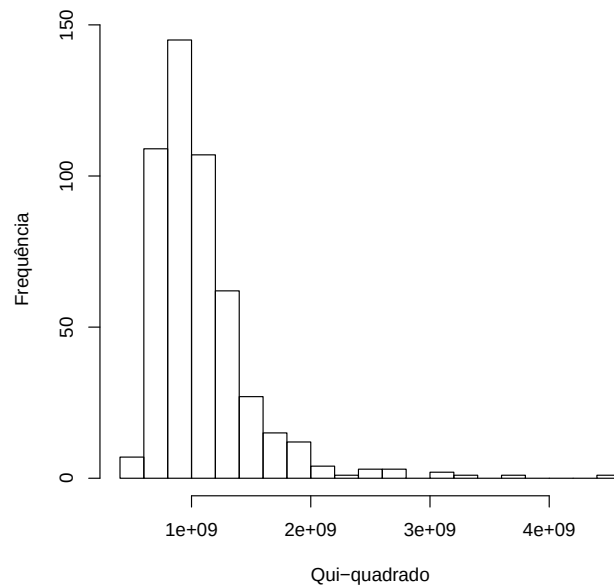


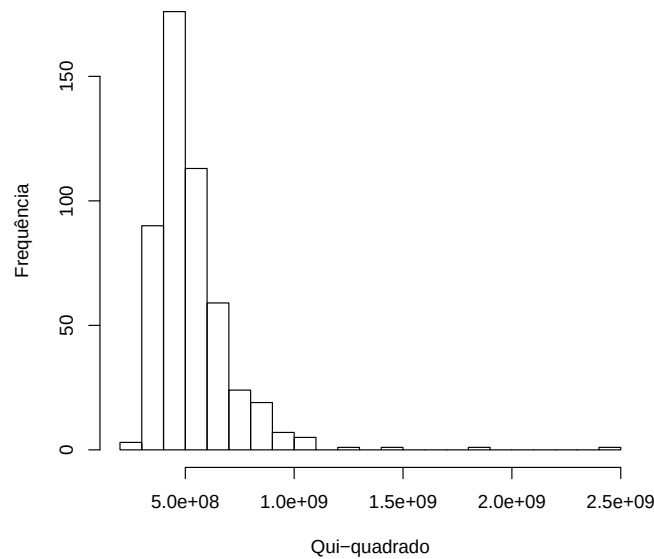
Figura 16: Estatísticas de teste do χ^2 - Modelo 1PL (2010)

No caso do modelo 2PL o percentil 95 da distribuição da estatística do teste do qui-quadrado gerado por *bootstrap* foi de 1769783448 e nos dados originais foi de 1416324471 (Tabela 11 e Figura 17).

p_0	p_{25}	p_{50}	p_{75}	p_{95}	p_{100}
459451646	772808639	917636798	1159312588	1769783448	5714615667

Tabela 11: Percentis do χ^2 - Modelo 2PL (2010)

Realizado o teste da razão de verossimilhanças para comparação da qualidade do ajustamento dos modelos 1PL e 2PL, concluiu-se pela não rejeição da hipótese nula, pelo que não existe evidência estatística de que o modelo 2PL se ajuste significativamente melhor aos dados do que o modelo 1PL.

Figura 17: Estatísticas de teste do χ^2 - Modelo 2PL (2010)

O percentil 95 da distribuição da estatística do teste da razão de verossimilhanças gerado por *bootstrap* foi de 114,5 sendo que nos dados originais foi de 56,5 (Tabela 12 e Figura 18).

p_0	p_{25}	p_{50}	p_{75}	p_{95}	p_{100}
43,8	71,6	84,0	97,1	114,5	158,6

Tabela 12: Percentis das razões de verossimilhança (2010)

A qualidade do ajustamento do modelo e a hipótese de independência local foram avaliadas analisando os resíduos marginais de ordem 2.

No caso do modelo 1PL detetaram-se diferenças significativas (resíduos qui-quadrado maiores que 3,5) entre as frequências observada e esperada em alguns padrões de resposta de pares de itens, a saber: (item 3, item 15), (item 4, item 15), (item 14, item 23) e (item 26, item 29), Tabela 13.

Pares de itens	(0,0)		(1,0)		(0,1)		(1,1)	
	O	E	O	E	O	E	O	E
(3,15)	-	-	-	-	49	37	28	40
(4,15)	-	-	46	35	-	-	16	28
(14,23)	56	43	-	-	27	40	-	-
(26,29)	32	20	22	34	-	-	-	-

Tabela 13: Avaliação marginal da qualidade do ajustamento - Modelo 1PL (2010)

A avaliação marginal da qualidade do ajustamento do modelo revela, portanto, qualidade insatisfatória, não se verificando a hipótese de independência local.

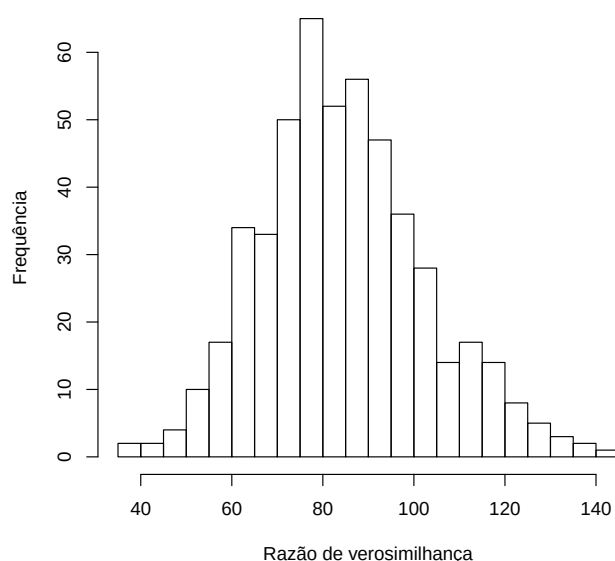


Figura 18: Estatísticas de teste da razão de verossimilhanças (2010)

Após ajustamento do modelo 2PL, não se detetaram diferenças significativas entre as frequências observada e esperada dos padrões de resposta para os diferentes pares de itens. A avaliação marginal da qualidade do ajustamento do modelo revela, portanto, boa qualidade, assumindo-se como verificada a hipótese de independência local.

Pelo exposto, neste estudo, retivemos o modelo 2PL acima descrito.

O histograma das habilidades previstas é apresentado na Figura 19 e as estimativas obtidas para os parâmetros de dificuldade e discriminação são apresentadas na Tabela 9.

Os parâmetros de dificuldade variaram entre -0,982 e 3,288, apresentando uma mediana de 0,225; os parâmetros de discriminação variaram entre 0,229 e 1,804, com uma mediana de 0,875. Relativamente aos parâmetros de discriminação o percentil 25 foi já de 0,769.

As cargas fatoriais estimadas variaram entre 0,223 e 0,875, com uma mediana de 0,658 e percentil 25 de 0,610. As curvas características dos itens são apresentadas na Figura 20. É possível observar que o item 11, relativo à resolução de uma inequação racional, e o item 15 apresentaram elevado parâmetro de dificuldade, uma vez que as respetivas curvas características dos itens se encontram deslocadas para a direita. O item 15, assim como o item 25, relacionado com o gráfico de uma função dada a sua derivada, apresentaram baixa capacidade de discriminação. A combinação dos elementos apresentados na Tabela 9 e Figura 20 sugere que o item 15 poderia ser eliminado da escala de avaliação final, por apresentar uma carga factorial baixa e ser pouco discriminativo. As respostas obtidas por categoria de opção para este item são apresentadas na Tabela 14, verificando-se que houve duas opções muito disputadas pelos alunos.

Tendo também em consideração o que se obteve para o ano 2009, verificou-se que os conteúdos cujos itens apresentaram maior parâmetro de dificuldade envolvem funções irracionais (inequações e derivadas), cálculo de derivadas (funções irracionais e exponenciais) e de limites com funções logarítmicas. Os itens mais acessíveis abordam inequações com

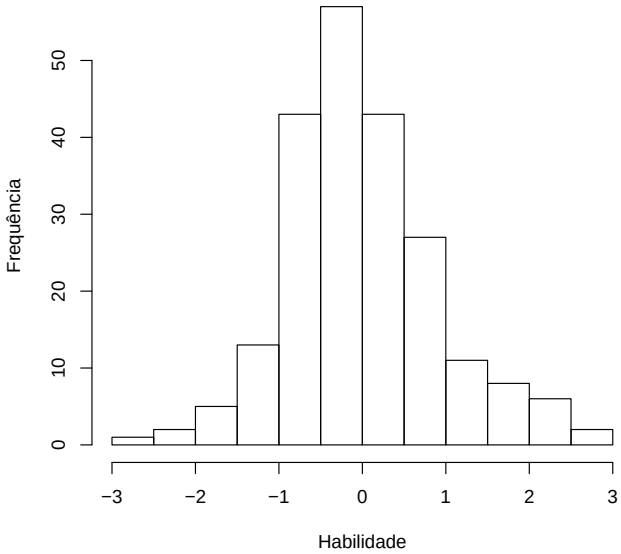


Figura 19: Habilidades previstas pelo modelo da TRI (2010)

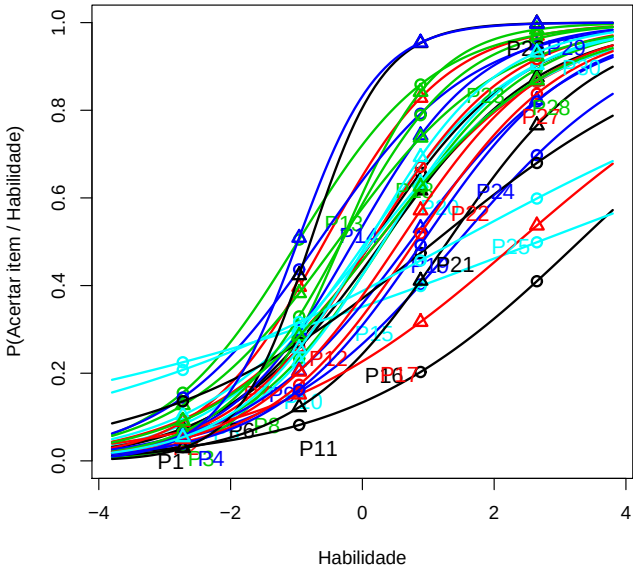


Figura 20: Curvas características dos itens (2010)

Item	NR	A	B	C	D
15	15 (7%)	77 (35%)	19 (9%)	92 (42%)	16 (4%)

Tabela 14: Respostas ao item 15 por opção (A, B, C ou D), para o ano de 2010. A célula a verde representa a opção correta.

módulos, propriedades dos logaritmos, função composta e o gráfico da derivada de uma função. Os conteúdos que mais discriminam os indivíduos estão relacionados com funções irracionais, função composta e com o gráfico da derivada de uma função. O conteúdo que se revelou menos discriminativo diz respeito ao cálculo da derivada de uma função irracional.

A curva de informação total do teste é apresentada na Figura 21.

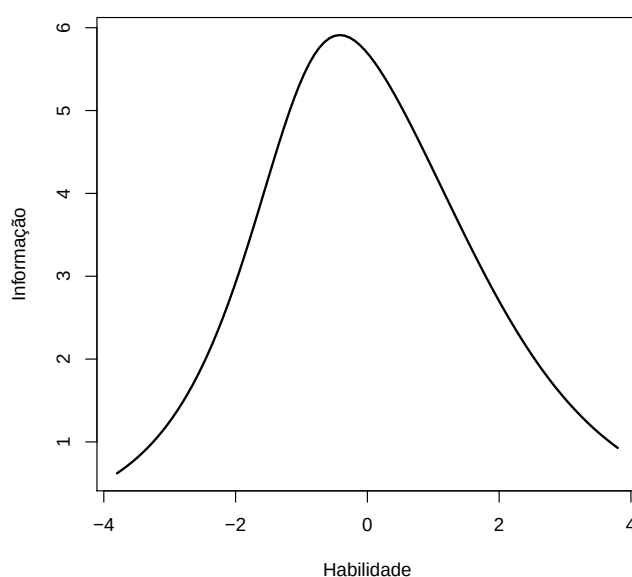


Figura 21: Curva de informação total (2010)

O gráfico do erro padrão de medição em função da habilidade corresponde à Figura 22. O ponto máximo de informação (5,911) foi atingido em 0,41 desvios padrão abaixo da média da habilidade. Este valor corresponde a um coeficiente alfa de Cronbach superior ao valor (constante) obtido da TCT. A partir da curva de informação, podem-se também retirar intervalos de confiança para a habilidade de cada aluno, com erro de medição dependente do padrão de resposta do aluno.

Relativamente à unidimensionalidade da habilidade, realizou-se uma análise em componentes principais sobre a matriz das correlações tetracóricas dos vários itens. A percentagem da variância total explicada pela primeira componente foi de 22,3% enquanto que essa percentagem foi de 6,4% para a segunda componente (Figura 23).

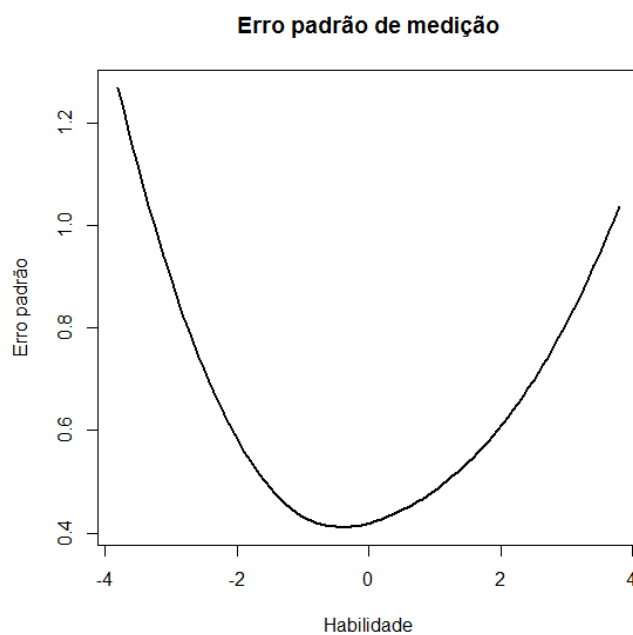


Figura 22: Erro padrão de medição (2010)

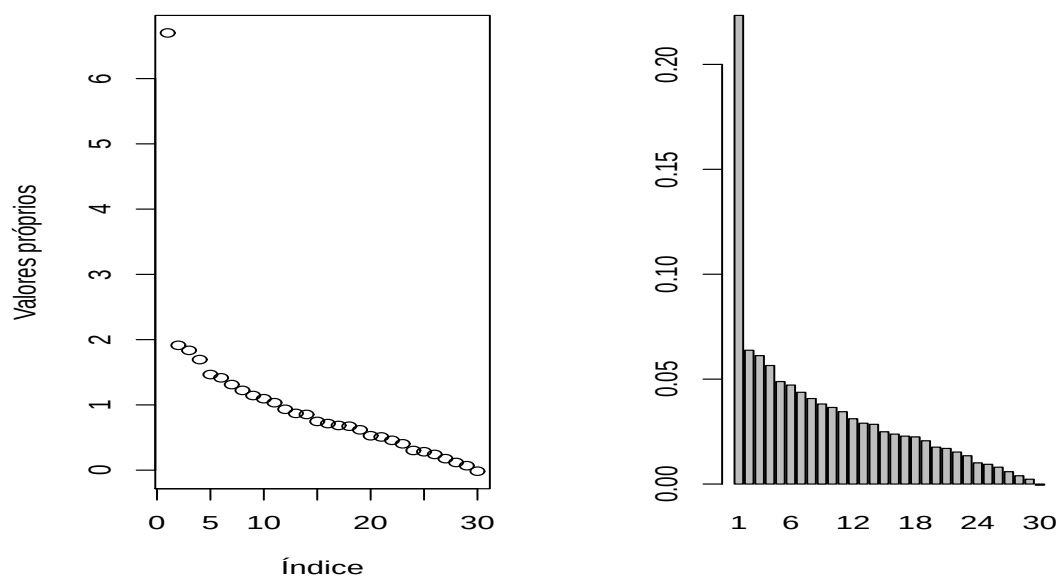


Figura 23: Valores próprios e componentes principais (2010)

A estatística *item-fit* mostrou que apenas os itens 6 e 15 não se ajustaram corretamente ao modelo seguido, enquanto que a estatística *person-fit* identificou apenas 4 padrões de resposta como inconsistentes para o modelo estimado.

3.3 Ano 2011

3.3.1 Descrição dos dados/Análise descritiva e exploratória

No ano 2011 o teste diagnóstico foi realizado no computador usando o programa Mooshak da autoria do Prof José Paulo Leal do Departamento de Ciência de Computadores da FCUP.

O teste foi constituído por 30 questões organizadas em 24 grupos. Por cada resposta correta é atribuído 1 ponto e por cada resposta errada $-\frac{1}{3}$. Os alunos podem seleccionar mais do que uma opção de resposta, contudo para análise do conhecimento do aluno a resposta só foi considerada correta se foi assinalada apenas a opção certa. Esta foi também a interpretação feita pela Prof Paula Lomp quando analisou os resultados do teste diagnóstico.

Associado a 19 dos 24 grupos de questões existe um banco de questões e o programa selecciona 1, 2 ou 4 questões dentro de cada grupo, dependendo do grupo. Estas questões foram preparadas por vários membros do Departamento de Matemática e pela Prof Ana Paula Tomás do Departamento de Ciências de Computadores da FCUP, que aliás foi responsável pela elaboração de uma metodologia de geração de opções para perguntas de escolha múltipla. A implementação computacional da metodologia e integração no Mooshak foi da responsabilidade do Prof José Paulo Leal. Os restantes grupos (G9, G11, G15, G18 e G24) consistiram em questões fixas iguais para todos os estudantes.

As perguntas foram organizadas da forma indicada na Tabela 15.

Numa primeira análise às questões do banco de itens observou-se existir grande heterogeneidade nos conhecimentos inerentes à resolução de alguns dos itens presentes no mesmo grupo. Verificou-se, por exemplo, no primeiro grupo de itens, referente a operações com números racionais, existirem itens cuja resolução envolve apenas a adição de números racionais e outros que envolvem para além desta operação a multiplicação ou divisão de números racionais, sendo, portanto, mais complexos. Situação análoga ocorreu relativamente a outros grupos de itens. Para fazer face a este facto foi ponderada a possibilidade de se reagruparem os itens em novos grupos mais homogéneos, contudo tal não se revelou viável dado o reduzido número de respostas que se obteriam a vários itens.

Imediatamente após a submissão do teste, os estudantes têm acesso à classificação obtida e à correção do que fizeram, ficando a saber as questões que erraram e quais seriam as respostas corretas.

Para análise dos resultados obtidos, dado o reduzido número de respostas que se obteriam a vários itens, optou-se por dicotomizar a cotação atribuída em cada grupo de questões. Atribuiu-se 1 ponto caso o estudante tenha obtido pelo menos metade da cotação máxima desse grupo de questões e 0 pontos caso contrário.

O teste em causa foi realizado por 328 estudantes e a classificação final podia, portanto, variar entre 0 e 24 pontos. Para simplificação da apresentação das classificações, a nota final foi considerada na escala usual de 0-20 valores.

3.3.2 Resultados

Responderam corretamente a mais de metade dos grupos de itens 187 estudantes (57%). A distribuição observada para as classificações (0-20 valores) foi razoavelmente simétrica, com uma mediana de 10,0 valores, uma média de 10,6 valores, e quartis de 7,5 e 13,3 valores. A classificação máxima atribuída foi de 20 valores (a dois alunos) e a mínima foi de 0 valores

G1 - Operações com números racionais	2 Questões
G2 - Operações com potências	2 Questões
G3 - Ordenação de números reais	1 Questão
G4 - Operações com radicais	1 Questão
G5 - Equações do primeiro grau	1 Questão
G6 - Equações do segundo ou terceiro grau	1 Questão
G7 - Conjunção e disjunção de inequações	1 Questão
G8 - Inequações do tipo $ax^2 + bx + c \geq d$ ou do tipo $\frac{ax+c}{bx+d} \geq 0$	1 Questão
G9 - " $ 3 + 2x \geq 1 \Leftrightarrow$ "	1 Questão
G10 - Domínios de expressões algébricas	1 Questão
G11 - "Se $x \geq 3$, $\sqrt{x^2 - 9} =$ "	1 Questão
G12 - Simplificação de expressões algébricas	1 Questão
G13 - Gráficos de funções (incluindo translações)	1 Questão
G14 - Identificação de retas pela equação	1 Questão
G15 - "Se $f(x) = \frac{1}{x}$ e $g(x) = e^x + 1$, então $(f \circ g)(x) =$ "	1 Questão
G16 - Divisão de polinômios	1 Questão
G17 - Senos e co-senos de alguns ângulos	1 Questão
G18 - "Se α é um ângulo do 4º quadrante e $\cos(\alpha) = \frac{1}{3}$ então $\sin(\alpha) =$ "	1 Questão
G19 - Utilização da fórmula fundamental da trigonometria	1 Questão
G20 - Cálculos básicos envolvendo logaritmos	4 Questões
G21 - Inequações com logaritmos e exponenciais	2 Questões
G22 - Limites de funções com $x \rightarrow \infty$ ou $x \rightarrow -\infty$	1 Questão
G23 - Limites de funções do tipo $\frac{p(x)}{x-a}$ quando $x \rightarrow a$	1 Questão
G24 - " $\lim_{x \rightarrow 2} \frac{x^2-4}{x^2+x-6}$ "	1 Questão

Tabela 15: Organização dos itens (2011)

(obtida por um aluno), Figura 24. Não se identificaram *outliers*.

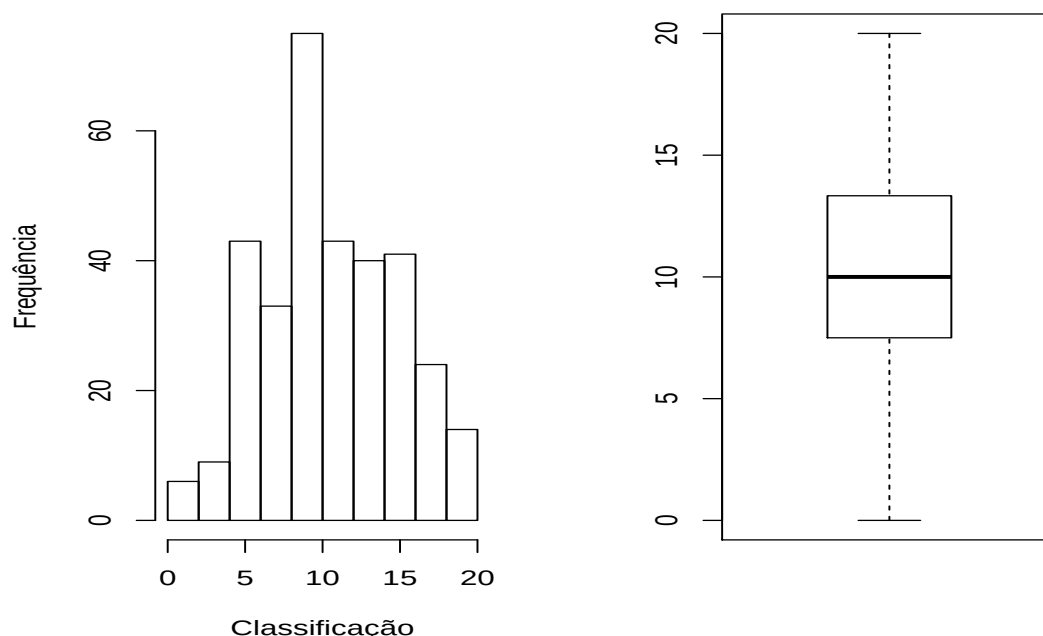


Figura 24: Classificações obtidas (2011)

3.3.2.1 Teoria Clássica de Testes (TCT)

Em relação às percentagens de respostas corretas por grupo de itens, a mediana foi de 50%, a média de 53%, os quartis de 46% e 60%, o mínimo foi de 22%, atingido no grupo 21, e o máximo foi de 90%, atingido no grupo 5 que é um *outlier* (Figura 25).

Quanto aos coeficientes de correlação bisserial entre cada um dos grupos de itens e a classificação final excluindo o grupo em causa, o valor mínimo encontrado foi de 0,311, no grupo de itens 1, e o máximo foi de 0,717, observado no grupo de questões 20. A média destas várias correlações foi de 0,466, a mediana foi de 0,482 e os quartis de 0,437 e 0,507. Todos os valores encontrados estão dentro do que é usual encontrar-se na teoria de testes para avaliação educacional (Tabela 16).

A consistência interna do teste foi aferida através do coeficiente alfa de Cronbach, avaliado em 0,825, e do erro padrão de medição, avaliado em 2,14. A amplitude dos intervalos a 95% de confiança para a classificação verdadeira foi de 7 valores, numa escala de 0 a 20 valores.

A avaliação da possível existência de grupos de itens não consistentes foi feita, para cada grupo de itens, através do cálculo do coeficiente alfa de Cronbach para o conjunto de grupos de itens restantes após o grupo em causa ter sido eliminado. Nenhuma eliminação de grupos de itens conduziu a um aumento do valor do coeficiente α de Cronbach correspondente à totalidade das perguntas (Tabela 16).

Item	TCT			TRI		
	p(%)	Correlação bisserial	α eliminando o item	Parâmetro de dificuldade	Parâmetro de discriminação	Fator de carga
G1	76	0,311	0,825	-1,896	0,671	0,557
G2	38	0,504	0,818	0,549	1,113	0,750
G3	50	0,441	0,820	0,015	0,871	0,657
G4	69	0,452	0,820	-0,951	1,007	0,710
G5	90	0,499	0,822	-2,145	1,253	0,782
G6	73	0,528	0,818	-1,002	1,255	0,782
G7	62	0,347	0,823	-0,810	0,677	0,561
G8	49	0,498	0,818	0,038	1,103	0,741
G9	49	0,554	0,816	0,032	1,294	0,791
G10	51	0,425	0,821	-0,033	0,893	0,666
G11	47	0,453	0,820	0,164	0,971	0,697
G12	43	0,470	0,819	0,343	0,982	0,701
G13	59	0,467	0,819	-0,428	1,017	0,713
G14	50	0,582	0,815	-0,018	1,407	0,815
G15	58	0,408	0,821	-0,452	0,851	0,648
G16	52	0,360	0,823	-0,113	0,719	0,584
G17	60	0,504	0,818	-0,462	1,053	0,725
G18	40	0,493	0,818	0,489	1,049	0,724
G19	48	0,535	0,817	0,087	1,198	0,768
G20	53	0,717	0,810	-0,119	2,214	0,911
G21	22	0,505	0,820	1,377	1,149	0,754
G22	64	0,510	0,818	-0,632	1,153	0,756
G23	33	0,365	0,823	1,066	0,731	0,590
G24	34	0,451	0,820	0,839	0,947	0,687
Média	53	0,466	0,819	-0,169	1,067	0,711

Tabela 16: Parâmetros estimados usando Teoria Clássica de Testes e Teoria de Resposta ao Item - Modelo 2PL (2011)

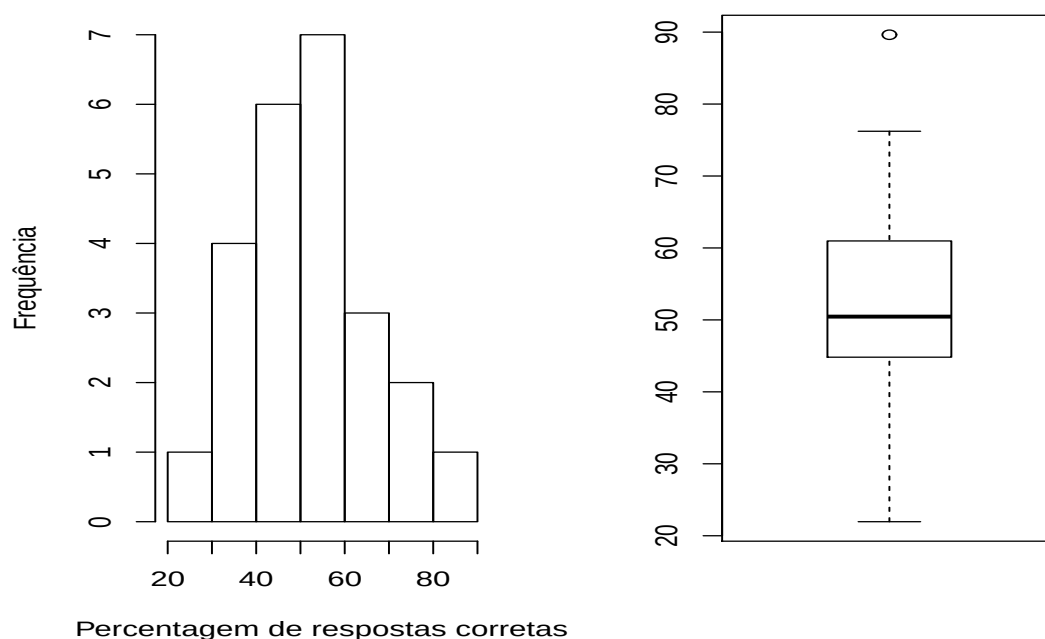


Figura 25: Percentagem de respostas corretas - para os itens (2011)

3.3.2.2 Teoria de Resposta ao Item (TRI)

Para análise da qualidade do ajustamento dos modelos logísticos com um parâmetro (1PL) e de dois parâmetros (2PL) procedeu-se de modo semelhante ao realizado para os anos anteriores.

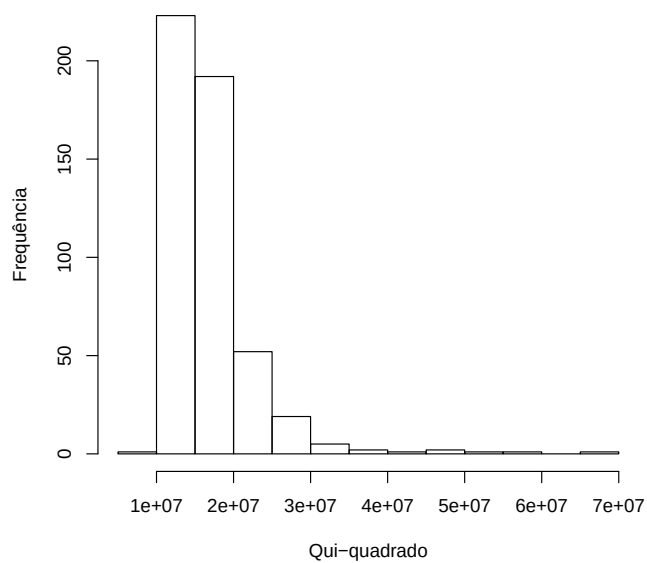
Realizado o procedimento anteriormente referido, concluiu-se pela rejeição da hipótese nula no caso do modelo logístico com um parâmetro (1PL), pelo que não revelou bom ajustamento aos dados, e pela não rejeição da hipótese nula no caso do modelo de dois parâmetros (2PL) pelo que este revelou bom ajustamento aos dados. Foram, portanto, identificadas diferenças estatisticamente significativas entre os dois modelos, no que concerne à qualidade do ajustamento, sendo melhor a do modelo de dois parâmetros (2PL).

No caso do modelo 1PL o percentil 95 da distribuição da estatística do teste do qui-quadrado gerado por *bootstrap* foi de 27392327 e nos dados originais foi de 28478539 (Tabela 17 e Figura 26).

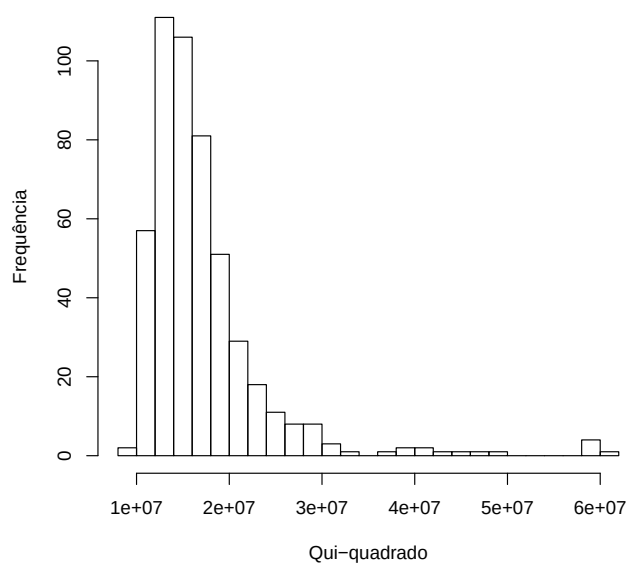
p_0	p_{25}	p_{50}	p_{75}	p_{95}	p_{100}
9165299	13454166	15411081	18599857	27392327	95088294

Tabela 17: Percentis do χ^2 - Modelo 1PL (2011)

No caso do modelo 2PL o percentil 95 da distribuição da estatística do teste do qui-quadrado gerado por *bootstrap* foi de 23968119 e nos dados originais foi de 28037997 (Tabela 18 e Figura 27).

Figura 26: Estatísticas de teste do χ^2 - Modelo 1PL (2011)

p_0	p_{25}	p_{50}	p_{75}	p_{95}	p_{100}
8697794	13109589	15166031	18578579	28037997	90815699

Tabela 18: Percentis do χ^2 - Modelo 2PL (2011)Figura 27: Estatísticas de teste do χ^2 - Modelo 2PL (2011)

Realizou-se o teste da razão de verossimilhanças para comparação da qualidade do ajustamento dos modelos 1PL e 2PL, concluindo-se pela não rejeição da hipótese nula, pelo que não existe evidência estatística de que o modelo 2PL se ajuste significativamente melhor aos dados do que o modelo 1PL.

O percentil 95 da distribuição da estatística do teste da razão de verossimilhanças gerado por *bootstrap* foi de 105,4 sendo que nos dados originais foi de 53,5 (Tabela 19 e Figura 28).

p_0	p_{25}	p_{50}	p_{75}	p_{95}	p_{100}
40,1	66,0	77,4	88,2	105,4	132,6

Tabela 19: Percentis das razões de verossimilhança (2011)

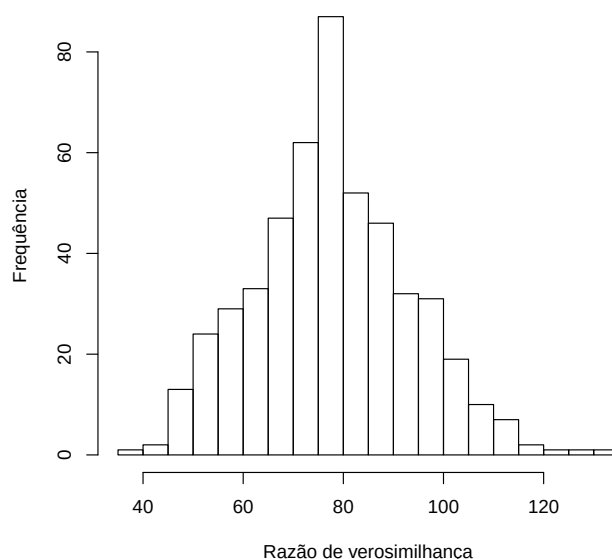


Figura 28: Estatísticas de teste da razão de verossimilhanças (2011)

A qualidade do ajustamento do modelo e a hipótese de independência local do modelo 2PL foram avaliadas analisando os resíduos marginais de ordem 2.

Após ajustamento do modelo, não se detetaram diferenças significativas (resíduos qui-quadrado maiores que 3,5) entre as frequências observada e esperada dos padrões de resposta para os diferentes pares de itens. A avaliação marginal da qualidade do ajustamento do modelo revela, portanto, boa qualidade, assumindo-se como verificada a hipótese de independência local.

Retivemos portanto, neste estudo, o modelo 2PL acima descrito.

O histograma das habilidades previstas é apresentado na Figura 29 e as estimativas obtidas para os parâmetros de dificuldade e discriminação são apresentadas na Tabela 16.

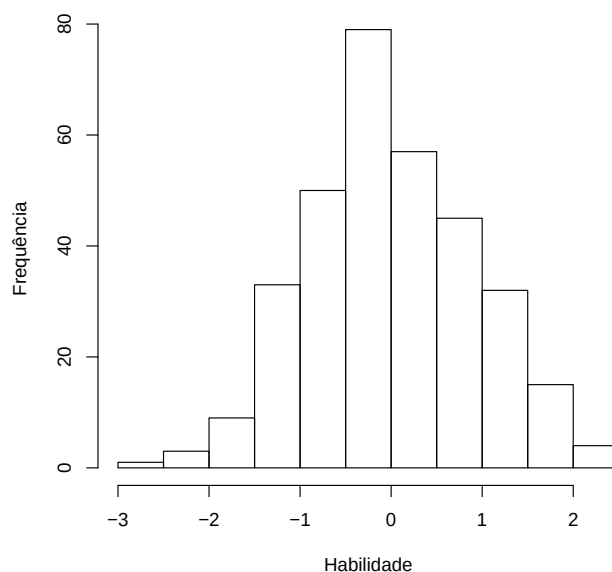


Figura 29: Habilidades previstas pelo modelo da TRI (2011)

Os parâmetros de dificuldade variaram entre -2,145 e 1,377, apresentando uma mediana de -0,026; os parâmetros de discriminação variaram entre 0,671 e 2,214, com uma mediana de 1,033. Relativamente aos parâmetros de discriminação o percentil 25 foi já de 0,889.

As cargas fatoriais estimadas variaram entre 0,557 e 0,911, com uma mediana de 0,719 e percentil 25 de 0,664. As curvas características dos itens são apresentadas na Figura 30. É possível observar que os itens G1 e G5, relativos a operações com números racionais e a equações do primeiro grau, respetivamente, apresentaram baixo parâmetro de dificuldade, uma vez que as respetivas curvas características dos itens se encontram deslocadas para a esquerda. A combinação dos elementos apresentados na Tabela 16 e Figura 30 sugere que todos os grupos de itens devem ser mantidos na escala de avaliação final.

A curva de informação total do teste é apresentada na Figura 31.

O gráfico do erro padrão de medição em função da habilidade corresponde à Figura 32. Nos valores da habilidade para os quais se tinha identificado um maior poder de discriminação dos itens a precisão é maior. O ponto máximo de informação (6,56) foi atingido em 0,112 desvios padrão abaixo da média da habilidade. Este valor corresponde a um coeficiente alfa de Cronbach superior ao valor (constante) obtido da TCT.

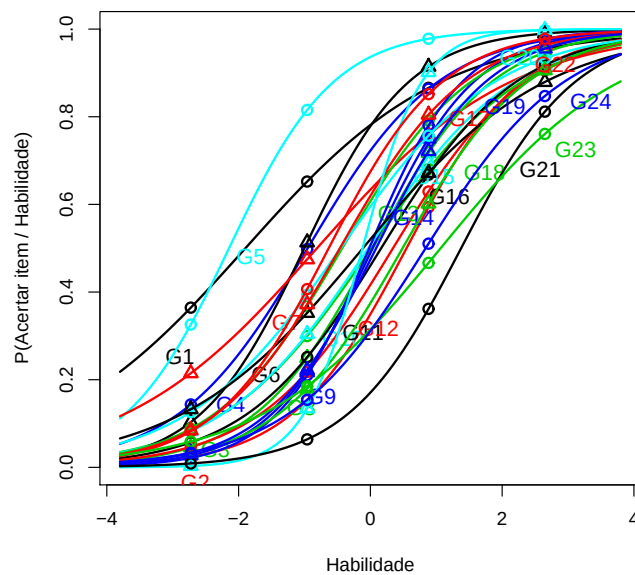


Figura 30: Curvas características dos itens (2011)

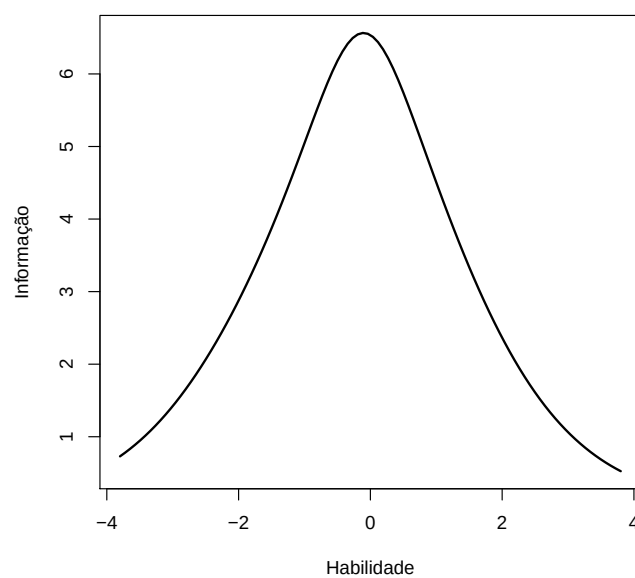


Figura 31: Curva de informação total (2011)

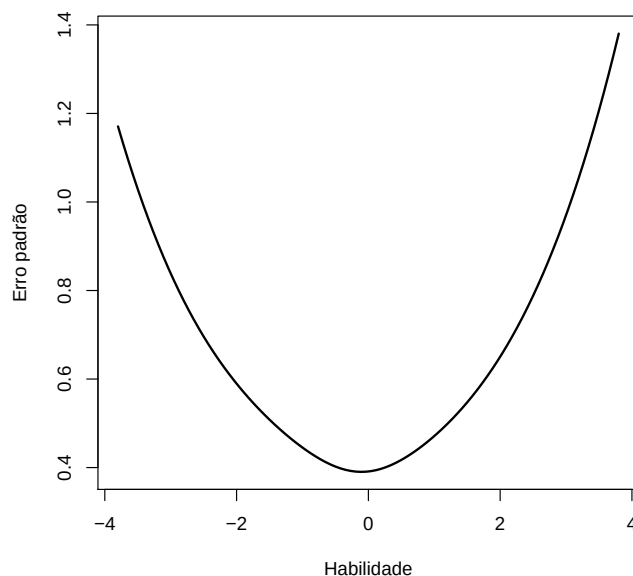


Figura 32: Erro padrão de medição (2011)

Relativamente à unidimensionalidade da habilidade esta foi comprovada através de uma análise em componentes principais sobre a matriz das correlações tetracóricas dos vários itens. A percentagem da variância total explicada pela primeira componente foi de 29,8% e de 6% para a segunda componente (Figura 33).

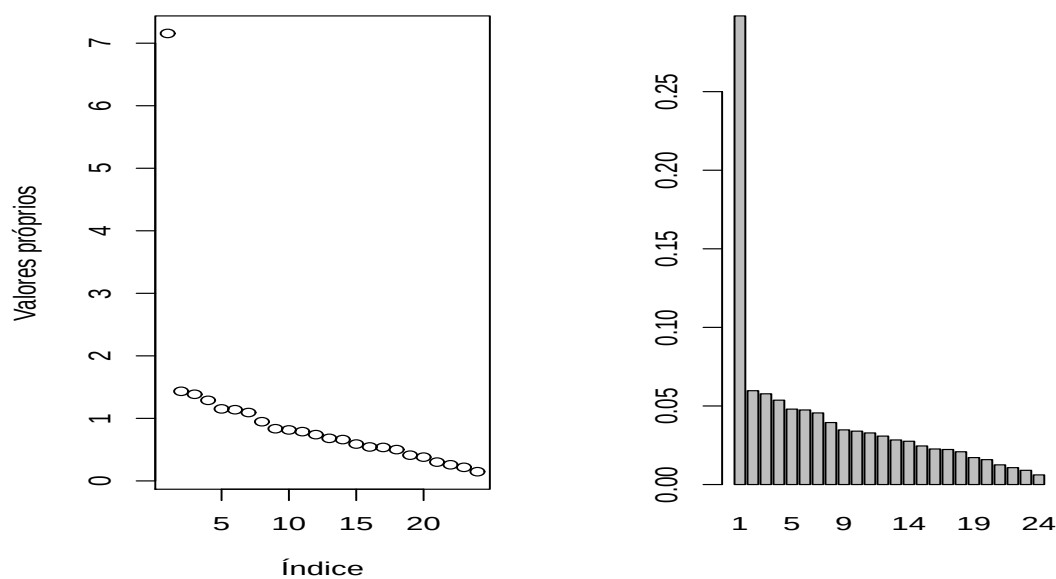


Figura 33: Valores próprios e componentes principais (2011)

A estatística *item-fit* mostrou que apenas os grupos de itens G8 e G16 não se ajustaram corretamente ao modelo seguido. A estatística *person-fit* identificou apenas 5 padrões de resposta como inconsistentes para o modelo estimado.

3.4 Ano 2012

3.4.1 Descrição dos dados/Análise descritiva e exploratória

No ano 2012 o teste diagnóstico foi realizado nos mesmos moldes do ano anterior. Realizaram o teste 206 estudantes tendo-se seguido a mesma abordagem na análise dos resultados obtidos.

3.4.2 Resultados

Dos estudantes que realizaram o teste 139 (67%) responderam corretamente a mais de metade dos grupos de itens. A distribuição observada para as classificações (0-20 valores) foi razoavelmente simétrica, com uma mediana de 12,1 valores, uma média também de 12,1 valores, e quartis de 9,2 e 15,8 valores. A classificação máxima atribuída foi de 20 valores (a um aluno) e a mínima foi de 2,5 valores (obtida por um aluno), Figura 34. Não se identificaram *outliers*.

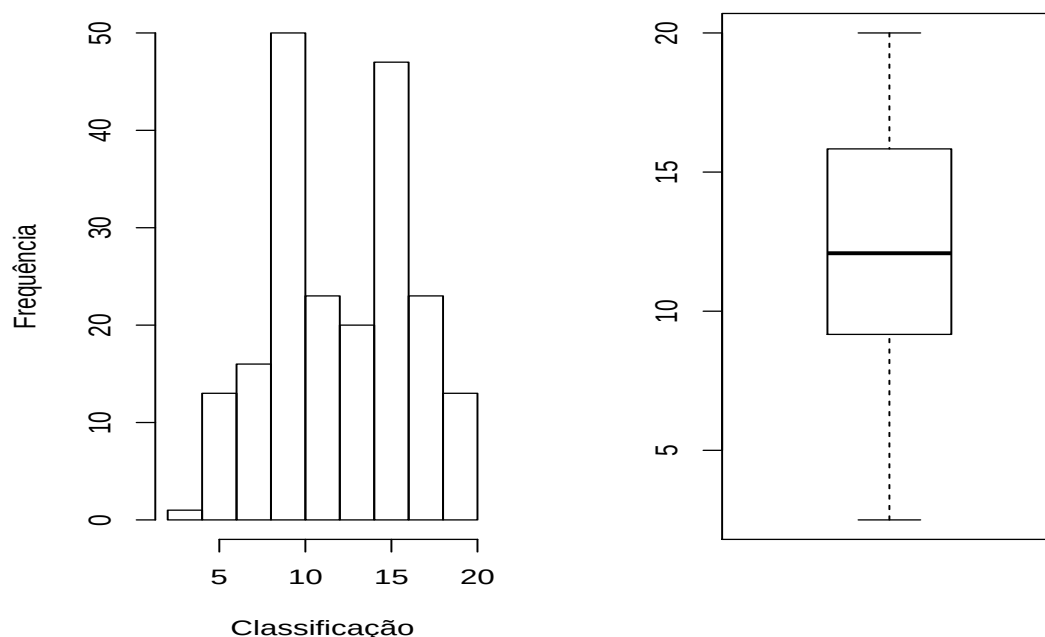


Figura 34: Classificações obtidas (2012)

3.4.2.1 Teoria Clássica de Testes (TCT)

Em relação às percentagens de respostas corretas por grupo de itens, a mediana foi de 61%, assim como a média, os quartis de 49% e 70%, o mínimo foi de 37%, atingido no grupo 21, e o máximo foi de 93%, atingido no grupo 5. Não se identificaram *outliers* (Figura 35).

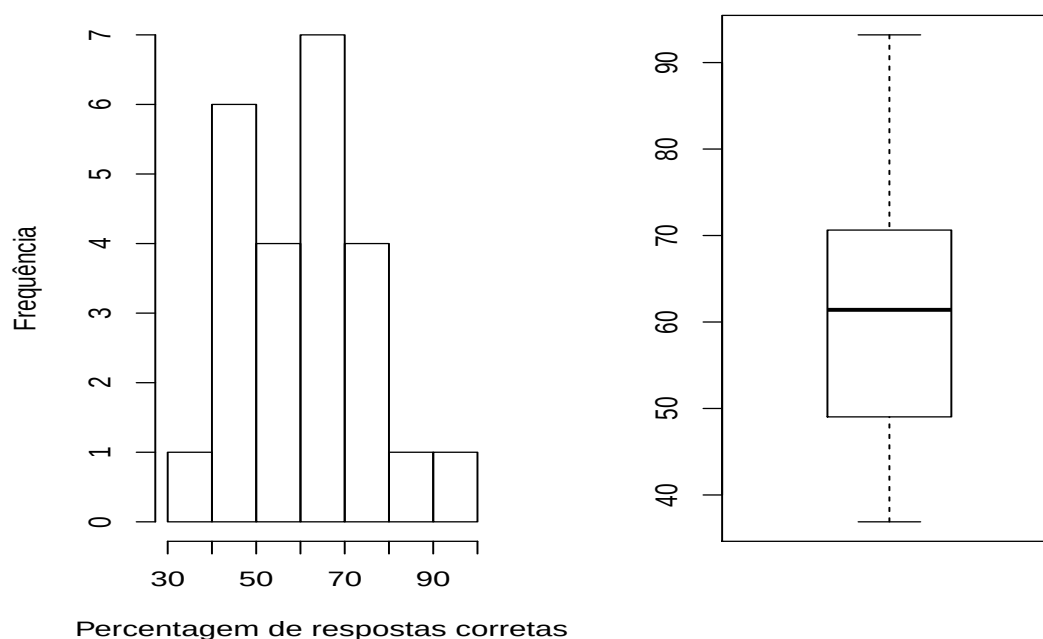


Figura 35: Percentagem de respostas corretas - para os itens (2012)

Quanto aos coeficientes de correlação bisserial entre cada um dos grupos de itens e a classificação final excluindo o grupo em causa, o valor mínimo encontrado foi de 0,187, no grupo de questões 5, e o máximo foi de 0,653, observado no grupo de questões 20. A média destas várias correlações foi de 0,419, a mediana foi de 0,446 e os quartis de 0,363 e 0,508. Com exceção do coeficiente mínimo os valores encontrados estão dentro do que é usual encontrar-se na teoria de testes para avaliação educacional (Tabela 20).

A consistência interna do teste foi aferida através do coeficiente alfa de Cronbach, avaliado em 0,797, e do erro padrão de medição, avaliado em 2,14. A amplitude dos intervalos a 95% de confiança para a classificação verdadeira foi de 7 valores, numa escala de 0 a 20 valores.

A avaliação da possível existência de grupos de itens não consistentes foi feita, para cada grupo de itens, através do cálculo do coeficiente alfa de Cronbach para o conjunto de grupos de itens restantes após o grupo em causa ter sido eliminado. Apenas a eliminação do grupo de itens 5 conduziu a um ligeiro aumento do valor do coeficiente alfa de Cronbach correspondente à totalidade das perguntas (Tabela 20). Como este item é também pouco discriminativo deve ser eliminado.

Item	TCT			TRI		
	p(%)	Correlação bisserial	α eliminando o item	Parâmetro de dificuldade	Parâmetro de discriminação	Fator de carga
G1	84	0,526	0,825	-1,591	1,331	0,800
G2	44	0,454	0,818	0,301	1,007	0,709
G3	55	0,303	0,820	-0,373	0,634	0,535
G4	71	0,334	0,820	-1,484	0,678	0,561
G5	93	0,187	0,822	-5,364	0,508	0,453
G6	74	0,431	0,818	-1,316	0,926	0,680
G7	70	0,372	0,823	-1,211	0,791	0,620
G8	63	0,542	0,818	-0,553	1,234	0,777
G9	65	0,341	0,816	-1,036	0,660	0,551
G10	57	0,455	0,821	-0,350	0,949	0,688
G11	48	0,454	0,820	0,115	0,963	0,693
G12	52	0,395	0,819	-0,116	0,799	0,624
G13	65	0,297	0,819	-1,091	0,620	0,527
G14	57	0,500	0,815	-0,344	1,085	0,735
G15	65	0,371	0,821	-0,920	0,765	0,608
G16	60	0,439	0,823	-0,530	0,937	0,684
G17	71	0,280	0,818	-1,709	0,574	0,498
G18	50	0,543	0,818	0,013	1,165	0,759
G19	49	0,521	0,817	0,058	1,132	0,749
G20	67	0,653	0,810	-0,609	1,786	0,873
G21	37	0,505	0,820	0,630	1,034	0,719
G22	74	0,458	0,818	-1,252	0,994	0,705
G23	46	0,517	0,823	0,181	1,205	0,769
G24	41	0,438	0,820	0,447	0,926	0,680
Média	61	0,419	0,819	-0,754	0,946	0,667

Tabela 20: Parâmetros estimados usando Teoria Clássica de Testes e Teoria de Resposta ao Item - Modelo 2PL (2012)

3.4.2.2 Teoria de Resposta ao Item (TRI)

Para análise da qualidade do ajustamento dos modelos logísticos com um parâmetro (1PL) e de dois parâmetros (2PL) procedeu-se de modo semelhante ao realizado para os anos anteriores.

Realizado o procedimento anteriormente referido, concluiu-se pela não rejeição da hipótese nula, pelo que quer o modelo logístico com um parâmetro (1PL) quer o de dois parâmetros (2PL) revelaram bom ajustamento aos dados. Não foram, portanto, identificadas diferenças estatisticamente significativas entre os dois modelos, no que concerne à qualidade do ajustamento.

No caso do modelo 1PL o percentil 95 da distribuição da estatística do teste do qui-quadrado gerado por *bootstrap* foi de 30305181 e nos dados originais foi de 18110222 (Tabela 21 e Figura 36).

p_0	p_{25}	p_{50}	p_{75}	p_{95}	p_{100}
7407512	11230940	13790866	17999147	30305181	124306266

Tabela 21: Percentis do χ^2 - Modelo 1PL (2012)

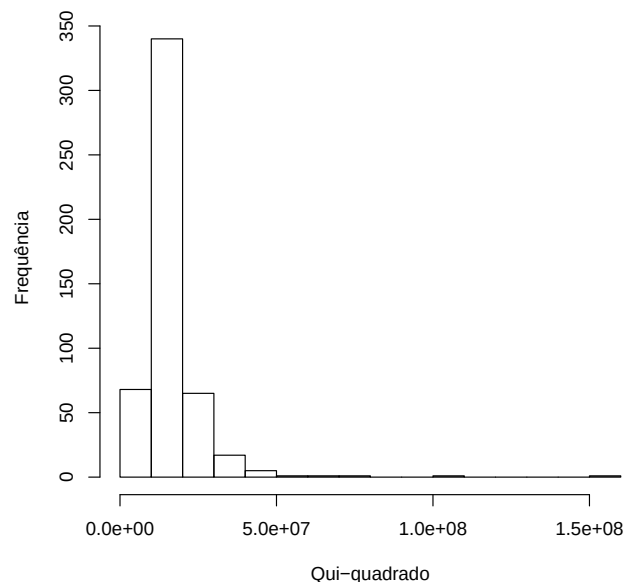
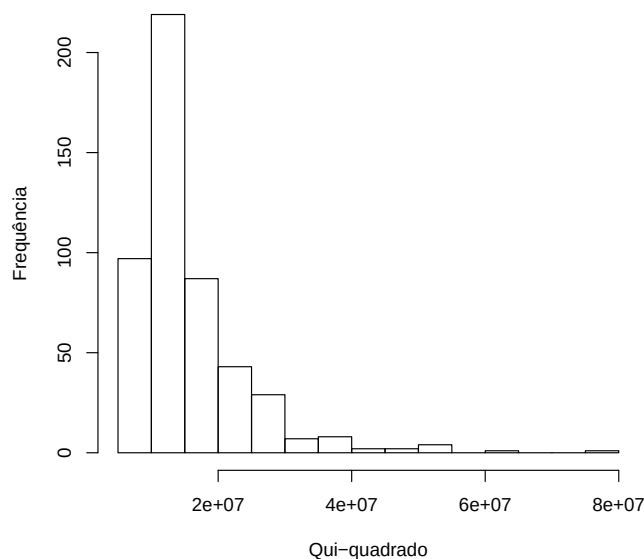


Figura 36: Estatísticas de teste do χ^2 - Modelo 1PL (2012)

No caso do modelo 2PL o percentil 95 da distribuição da estatística do teste do qui-quadrado gerado por *bootstrap* foi de 31966147 e nos dados originais foi de 20945012 (Tabela 22 e Figura 37).

p_0	p_{25}	p_{50}	p_{75}	p_{95}	p_{100}
6657950	10999707	13291249	17057877	31966147	98206397

Tabela 22: Percentis do χ^2 - Modelo 2PL (2012)Figura 37: Estatísticas de teste do χ^2 - Modelo 2PL (2012)

Realizado o teste da razão de verossimilhanças para comparação da qualidade do ajustamento dos modelos 1PL e 2PL, concluiu-se pela não rejeição da hipótese nula, pelo que não existe evidência estatística de que o modelo 2PL se ajuste significativamente melhor aos dados do que o modelo 1PL.

O percentil 95 da distribuição da estatística do teste da razão de verossimilhanças gerado por *bootstrap* foi de 81,1 sendo que nos dados originais foi de 33,4 (Tabela 23 e Figura 38).

p_0	p_{25}	p_{50}	p_{75}	p_{95}	p_{100}
27,9	48,8	56,4	67,2	81,1	113,5

Tabela 23: Percentis das razões de verossimilhança (2012)

A qualidade do ajustamento do modelo e a hipótese de independência local foram avaliadas analisando os resíduos marginais de ordem 2.

No caso do modelo 1PL detetaram-se diferenças significativas (resíduos qui-quadrado maiores que 3,5) entre as frequências observada e esperada em alguns padrões de resposta de pares de itens, a saber: (item 1, item 16) e (item 23, item 24), Tabela 24.

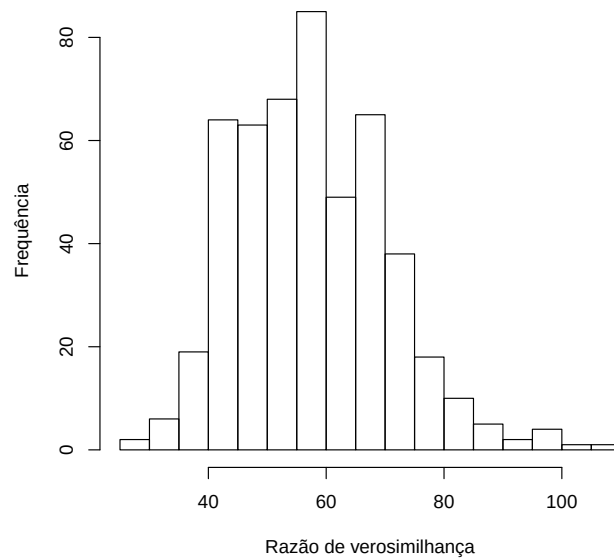


Figura 38: Estatísticas de teste da razão de verossimilhanças (2012)

Pares de itens	(1,0)		(0,1)		(1,1)	
	O	E	O	E	O	E
(1,16)	-	-	8	16	-	-
(23,24)	34	48	25	38	60	47

Tabela 24: Avaliação marginal da qualidade do ajustamento - Modelo 1PL (2012)

No caso do modelo 2PL detetaram-se diferenças significativas entre as frequências observada e esperada dos padrões de resposta para os pares de itens (item 17, item 22) e (item 23, item 24), Tabela 25.

Pares de itens	(0,0)		(0,1)	
	O	E	O	E
(17,22)	28	19	-	-
(23,24)	-	-	25	38

Tabela 25: Avaliação marginal da qualidade do ajustamento - Modelo 2PL (2012)

A avaliação marginal da qualidade do ajustamento do modelo revela, portanto, boa qualidade, assumindo-se como verificada a hipótese de independência local.

Após ajustamento do modelo 2PL, não se detetaram diferenças significativas entre as frequências observada e esperada dos padrões de resposta para os diferentes pares de itens. A avaliação marginal da qualidade do ajustamento do modelo revela, portanto, boa qualidade, assumindo-se como verificada a hipótese de independência local.

Pelo exposto, neste estudo, retivemos o modelo 2PL acima descrito.

O histograma das habilidades previstas é apresentado na Figura 39 e as estimativas obtidas para os parâmetros de dificuldade e discriminação são apresentadas na Tabela 20.

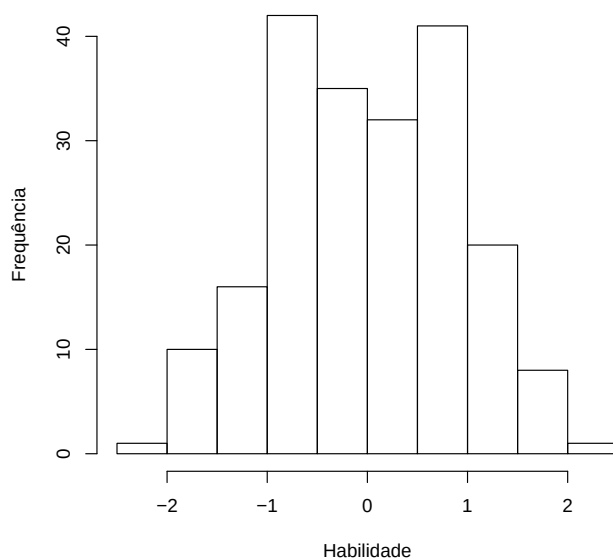


Figura 39: Habilidades previstas pelo modelo da TRI (2012)

Os parâmetros de dificuldade variaram entre -5,364 e 0,630, apresentando uma mediana de -0,542; os parâmetros de discriminação variaram entre 0,508 e 1,786, com uma mediana de 0,943. Relativamente aos parâmetros de discriminação o percentil 25 foi já de 0,743.

As cargas fatoriais estimadas variaram entre 0,453 e 0,873, com uma mediana de 0,686 e percentil 25 de 0,596. As curvas características dos itens são apresentadas na Figura 40. É possível observar, tal como no ano anterior, que os itens G1 e G5 apresentaram baixo parâmetro de dificuldade, no caso do item G1 muito baixo. Para além disso, o item G20, relativo a operações com logaritmos, registou um elevado parâmetro de discriminação. A combinação dos elementos apresentados na Tabela 20 e Figura 40 sugere que todos os grupos de itens devem ser mantidos na escala de avaliação final.

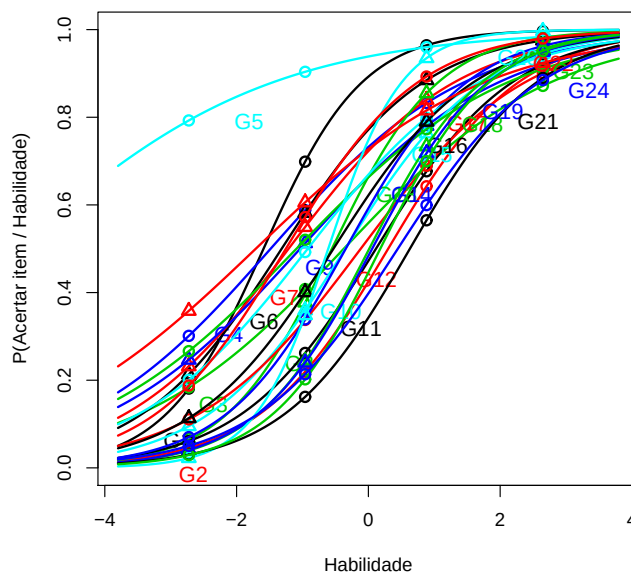


Figura 40: Curvas características dos itens (2012)

A curva de informação total do teste é apresentada na Figura 41.

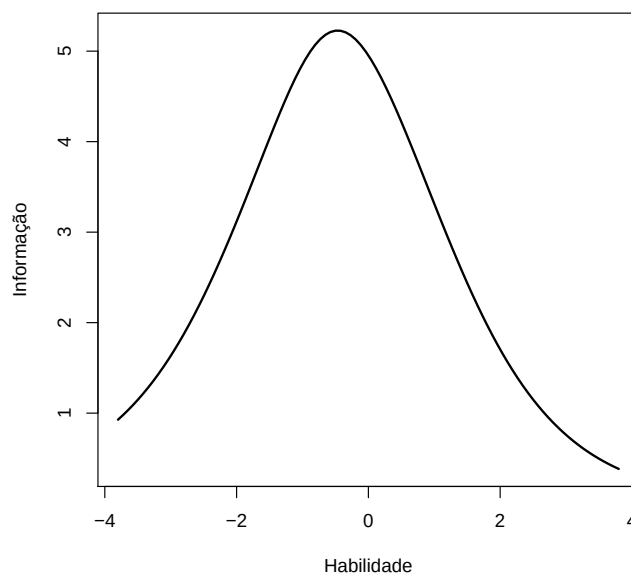


Figura 41: Curva de informação total (2012)

O gráfico do erro padrão de medição em função da habilidade é representado na Figura 42. O ponto máximo de informação (5,23) foi atingido em 0,468 desvios padrão abaixo da média da habilidade. Este valor corresponde a um coeficiente alfa de Cronbach superior ao

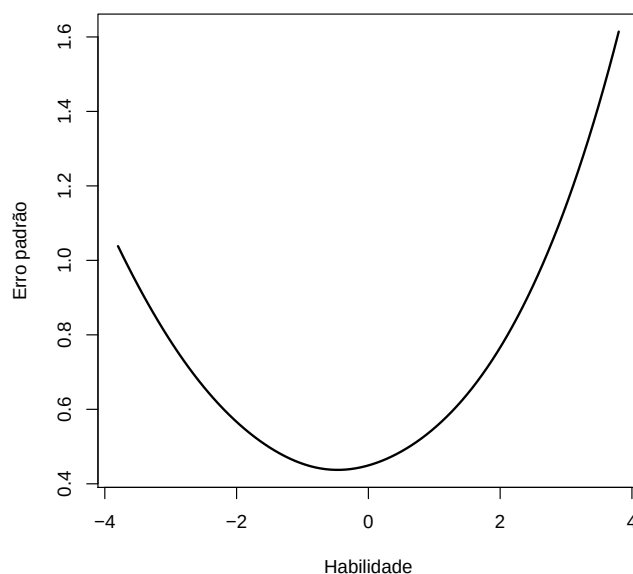


Figura 42: Erro padrão de medição (2012)

valor (constante) obtido da TCT.

No que concerne à unidimensionalidade da habilidade, realizou-se uma análise em componentes principais sobre a matriz das correlações tetracóricas dos vários itens. A percentagem da variância total explicada pela primeira componente foi de 26,3% enquanto que essa percentagem foi de 8,2% para a segunda componente (Figura 43).

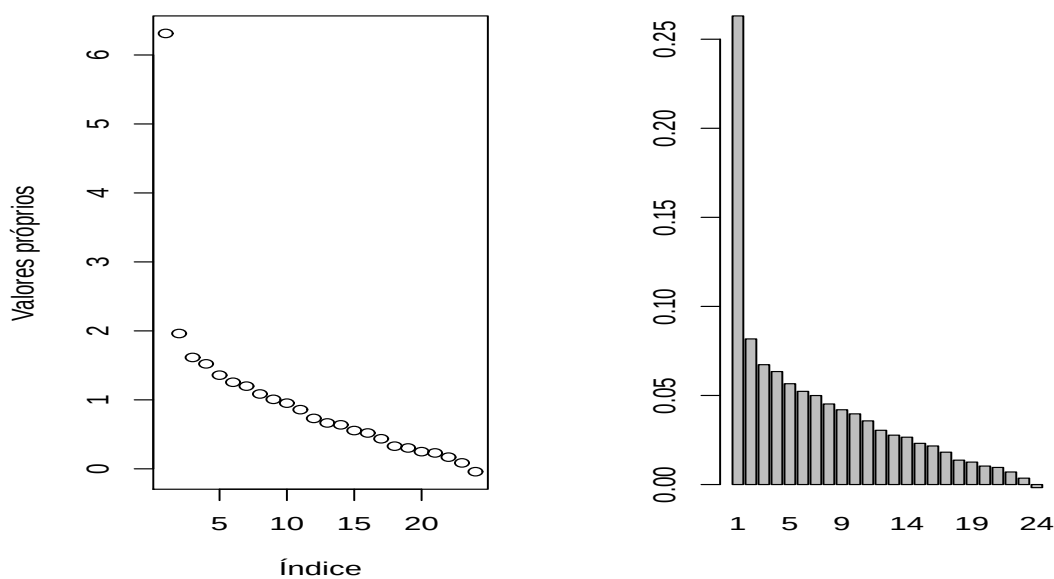


Figura 43: Valores próprios e componentes principais (2012)

A estatística *item-fit* mostrou que apenas os grupos de itens G2 e G10 não se ajustaram corretamente ao modelo seguido, enquanto que a estatística *person-fit* identificou apenas 6 padrões de resposta como inconsistentes para o modelo estimado.

3.5 Ano 2013

3.5.1 Descrição dos dados/Análise descritiva e exploratória

No ano 2013 o teste diagnóstico foi realizado nos mesmos moldes do ano anterior. Seguiu-se a mesma abordagem na análise dos resultados obtidos.

Participaram no teste apenas 150 estudantes dos quais 2 não conseguiram submeter o teste. Tendo-se verificado que em alguns cursos nenhum aluno tinha participado, foi dada uma nova possibilidade aos interessados de realizarem o teste, contudo nesta segunda fase apenas compareceram 16 alunos.

3.5.2 Resultados

Dos 164 estudantes que efetivamente realizaram o teste 105 (64%) responderam corretamente a mais de metade dos grupos de itens. A distribuição observada para as classificações (0-20 valores) foi razoavelmente simétrica, com uma mediana de 11,7 valores, uma média de 11,8 valores, e quartis de 8,3 e 15 valores. A classificação máxima atribuída foi de 20 valores (a quatro alunos) e a mínima foi de 0 valores (obtida por dois alunos), Figura 44. Não se identificaram *outliers*.

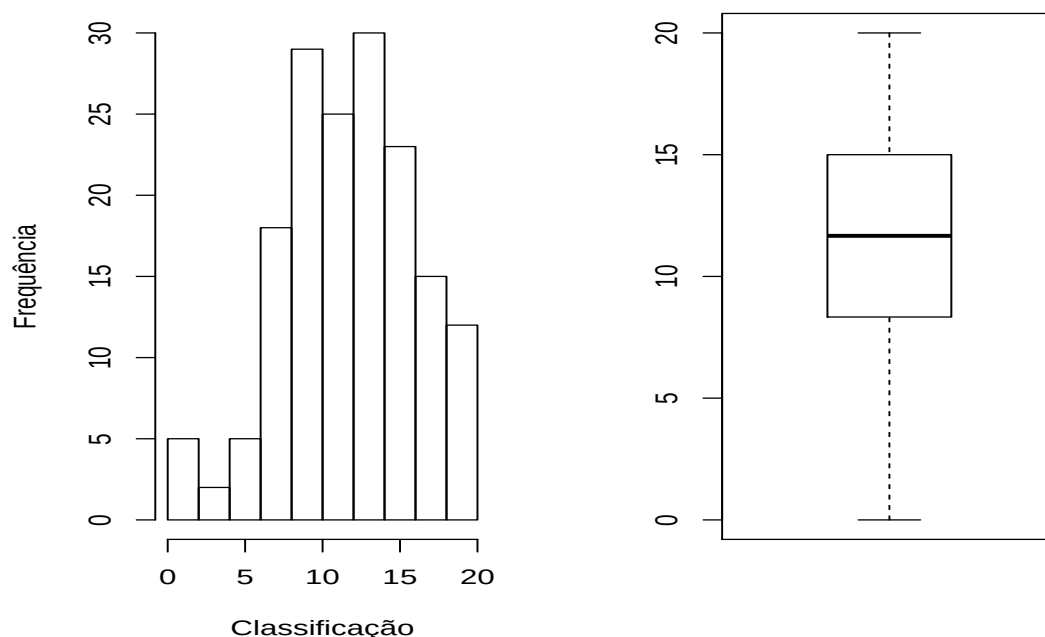


Figura 44: Classificações obtidas (2013)

3.5.2.1 Teoria Clássica de Testes (TCT)

Em relação às percentagens de respostas corretas por grupo de itens, a mediana foi de 57%, a média foi de 59%, os quartis de 48% e 70%, o mínimo foi de 31%, atingido no grupo 21, e o máximo foi de 93%, atingido no grupo 5. Não se identificaram *outliers* (Figura 45).

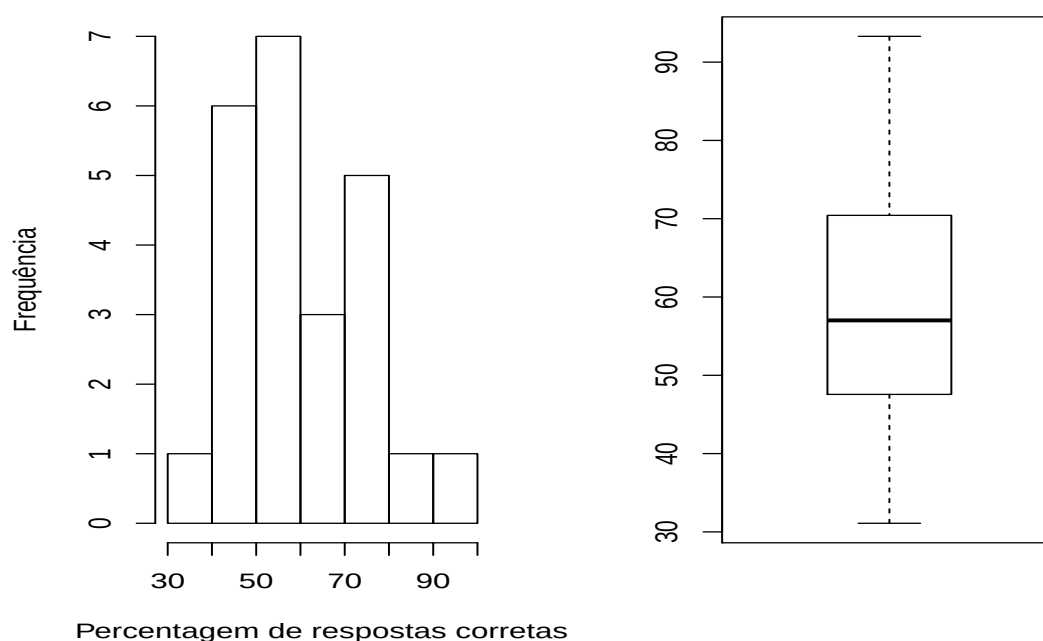


Figura 45: Percentagem de respostas corretas - para os itens (2013)

Quanto aos coeficientes de correlação bisserial entre cada um dos grupos de itens e a classificação final excluindo o grupo em causa, o valor mínimo encontrado foi de 0,328, no grupo de questões 3, e o máximo foi de 0,680, observado no grupo de questões 20. A média destas várias correlações foi de 0,495, a mediana foi de 0,501 e os quartis de 0,461 e 0,548. Todos os valores encontrados estão dentro do que é usual encontrar-se na teoria de testes para avaliação educacional (Tabela 26).

A consistência interna do teste foi aferida através do coeficiente alfa de Cronbach, avaliado em 0,835, e do erro padrão de medição, avaliado em 2,1. A amplitude dos intervalos a 95% de confiança para a classificação verdadeira foi de 6,9 valores, numa escala de 0 a 20 valores.

A avaliação da possível existência de grupos de itens não consistentes foi feita, para cada grupo de itens, através do cálculo do coeficiente alfa de Cronbach para o conjunto de grupos de itens restantes após o grupo em causa ter sido eliminado. Nenhuma eliminação de grupos de itens conduziu a um aumento do valor do coeficiente alfa de Cronbach correspondente à totalidade das perguntas (Tabela 26).

Item	TCT			TRI		
	p(%)	Correlação bisserial	α eliminando o item	Parâmetro de dificuldade	Parâmetro de discriminação	Fator de carga
G1	84	0,556	0,830	-1,620	1,306	0,794
G2	52	0,587	0,825	-0,070	1,590	0,846
G3	55	0,328	0,834	-0,325	0,655	0,548
G4	76	0,505	0,829	-1,210	1,147	0,754
G5	93	0,579	0,832	-2,444	1,360	0,806
G6	72	0,516	0,829	-0,985	1,190	0,766
G7	66	0,442	0,830	-0,873	0,907	0,672
G8	53	0,545	0,827	-0,122	1,344	0,802
G9	51	0,458	0,830	-0,029	1,021	0,714
G10	58	0,497	0,828	-0,336	1,203	0,769
G11	41	0,557	0,826	0,328	1,374	0,808
G12	46	0,466	0,829	0,192	1,082	0,734
G13	65	0,507	0,828	-0,635	1,188	0,765
G14	57	0,462	0,830	-0,311	1,047	0,723
G15	70	0,472	0,830	-0,949	1,091	0,737
G16	57	0,424	0,831	-0,378	0,906	0,671
G17	63	0,366	0,833	-0,824	0,741	0,596
G18	49	0,489	0,829	0,049	1,197	0,768
G19	46	0,484	0,829	0,157	1,147	0,754
G20	71	0,680	0,824	-0,661	2,183	0,909
G21	31	0,435	0,831	0,976	0,960	0,692
G22	72	0,618	0,826	-0,781	1,754	0,869
G23	43	0,535	0,827	0,265	1,309	0,795
G24	45	0,506	0,828	0,202	1,208	0,770
Média	59	0,495	0,829	-0,433	1,205	0,753

Tabela 26: Parâmetros estimados usando Teoria Clássica de Testes e Teoria de Resposta ao Item - Modelo 2PL (2013)

3.5.2.2 Teoria de Resposta ao Item (TRI)

Para análise da qualidade do ajustamento dos modelos logísticos com um parâmetro (1PL) e de dois parâmetros (2PL) procedeu-se de modo semelhante ao realizado para os anos anteriores.

Realizado o procedimento anteriormente referido, concluiu-se pela não rejeição da hipótese nula, pelo que quer o modelo logístico com um parâmetro (1PL) quer o de dois parâmetros (2PL) revelaram bom ajustamento aos dados. Não foram, portanto, identificadas diferenças estatisticamente significativas entre os dois modelos, no que concerne à qualidade do ajustamento.

No caso do modelo 1PL o percentil 95 da distribuição da estatística do teste do qui-quadrado gerado por *bootstrap* foi de 30069980 e nos dados originais foi de 9885059 (Tabela 27 e Figura 46).

p_0	p_{25}	p_{50}	p_{75}	p_{95}	p_{100}
6374958	10290551	12781295	17166294	30069980	105182995

Tabela 27: Percentis do χ^2 - Modelo 1PL (2013)

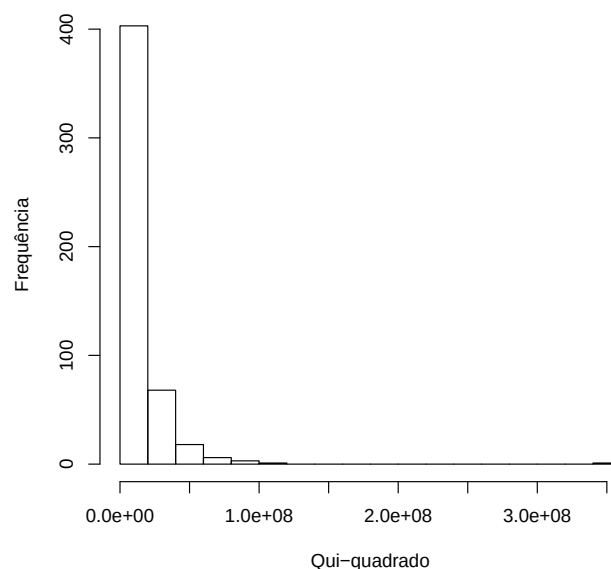
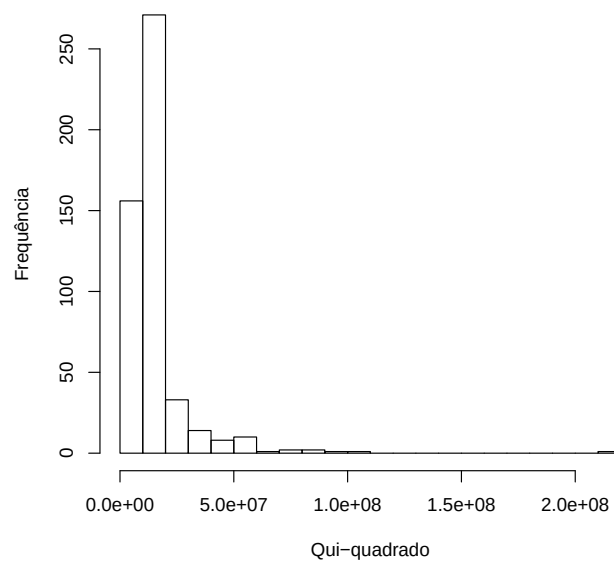


Figura 46: Estatísticas de teste do χ^2 - Modelo 1PL (2013)

No caso do modelo 2PL o percentil 95 da distribuição da estatística do teste do qui-quadrado gerado por *bootstrap* foi de 35661156 e nos dados originais foi de 10458913 (Tabela 28 e Figura 47).

p_0	p_{25}	p_{50}	p_{75}	p_{95}	p_{100}
3698647	8415147	11262235	15495786	35661156	519789846

Tabela 28: Percentis do χ^2 - Modelo 2PL (2013)Figura 47: Estatísticas de teste do χ^2 - Modelo 2PL (2013)

Realizou-se o teste da razão de verossimilhanças para comparação da qualidade do ajustamento dos modelos 1PL e 2PL, concluindo-se pela não rejeição da hipótese nula, pelo que não existe evidência estatística de que o modelo 2PL se ajuste significativamente melhor aos dados do que o modelo 1PL.

O percentil 95 da distribuição da estatística do teste da razão de verossimilhanças gerado por *bootstrap* foi de 68,9 sendo que nos dados originais foi de 23,6 (Tabela 29 e Figura 48).

p_0	p_{25}	p_{50}	p_{75}	p_{95}	p_{100}
17,7	37,4	46,2	54,6	68,9	101,1

Tabela 29: Percentis das razões de verossimilhança (2013)

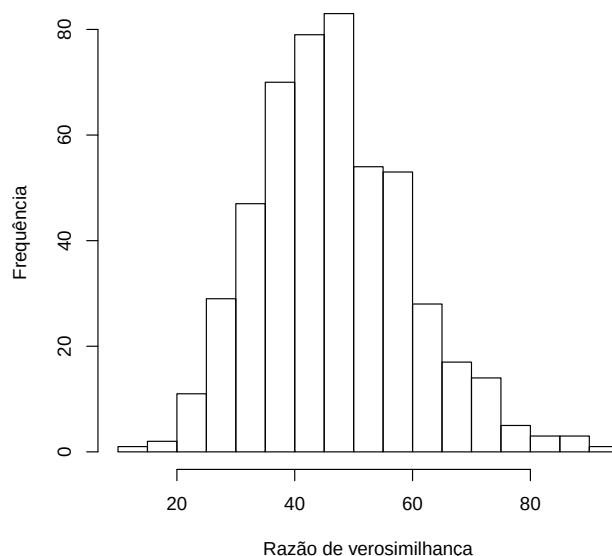


Figura 48: Estatísticas de teste da razão de verossimilhanças (2013)

Analizados os resíduos marginais de ordem 2 para estudo da hipótese de independência local e da qualidade do ajustamento do modelo observou-se que, no caso do modelo 1PL existem diferenças significativas (resíduos qui-quadrado maiores que 3,5) entre as frequências observada e esperada em alguns padrões de resposta de pares de itens, a saber: (item 23, item 24). Verificou-se que os alunos que acertaram o item 23 e erraram o item 24 o fizeram um número de vezes (20) inferior ao previsto pelo modelo (31).

Após ajustamento do modelo 2PL, não se detetaram diferenças significativas entre as frequências observada e esperada dos padrões de resposta para os diferentes pares de itens. Dada a avaliação marginal da qualidade do ajustamento do modelo assume-se como verificada a hipótese de independência local.

À semelhança do sucedido nos anos anteriores, neste estudo, retivemos o modelo 2PL acima descrito.

O histograma das habilidades previstas é apresentado na Figura 49 e as estimativas obtidas para os parâmetros de dificuldade e discriminação são apresentadas na Tabela 26.

Os parâmetros de dificuldade variaram entre -2,444 e 0,976, apresentando uma mediana de -0,330; os parâmetros de discriminação variaram entre 0,655 e 2,183, com uma mediana de 1,189. Relativamente aos parâmetros de discriminação o percentil 25 foi já de 1,040.

As cargas fatoriais estimadas variaram entre 0,548 e 0,909, com uma mediana de 0,765 e percentil 25 de 0,721. As curvas características dos itens são apresentadas na Figura 50. É possível observar, à semelhança do ano anterior, que os itens G1 e G5 apresentaram baixo parâmetro de dificuldade e que o item G20 registou um elevado parâmetro de discriminação. A combinação dos elementos apresentados na Tabela 26 e Figura 50 sugere que todos os grupos de itens devem ser mantidos na escala de avaliação final.

Tendo também em consideração o que se obteve para os anos 2011 e 2012, anos em que o teste diagnóstico também foi realizado no computador, com recurso a um banco

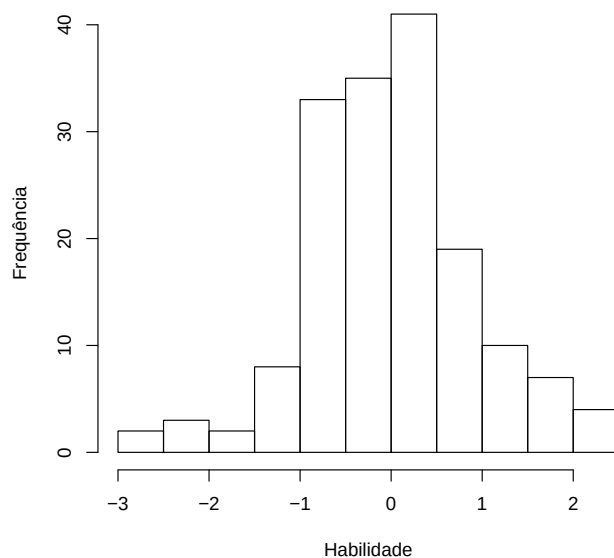


Figura 49: Habilidades previstas pelo modelo da TRI (2013)

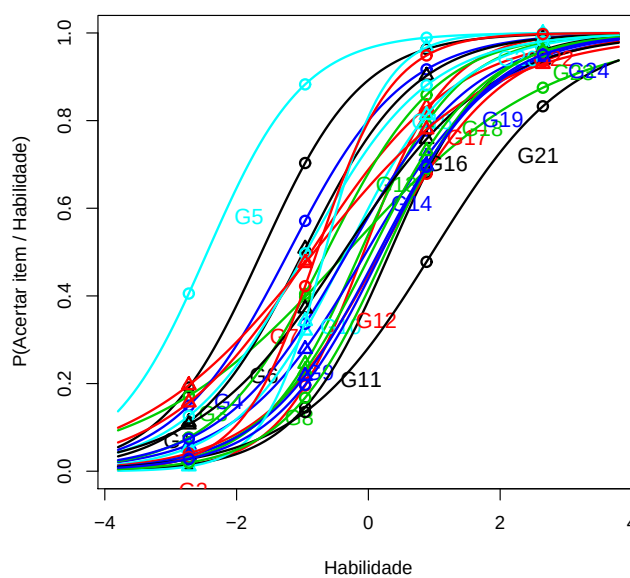


Figura 50: Curvas características dos itens (2013)

de questões, verificou-se que os conteúdos cujos itens apresentaram maior parâmetro de dificuldade envolvem inequações com logaritmos e exponenciais. Os itens mais acessíveis estão relacionados com operações com números racionais e equações do primeiro, segundo e terceiro grau. Os conteúdos que mais discriminam os indivíduos envolvem cálculos com logaritmos, não se tendo registado itens pouco discriminativos.

A curva de informação total do teste é apresentada na Figura 51.

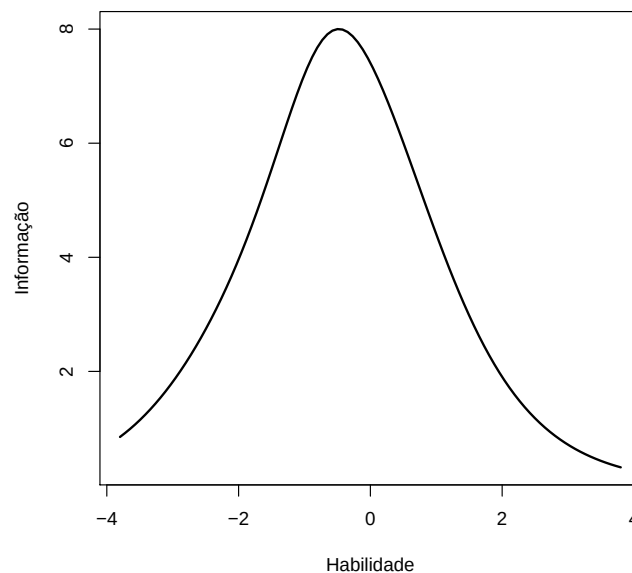


Figura 51: Curva de informação total (2013)

O gráfico do erro padrão de medição em função da habilidade corresponde à Figura 52. O ponto máximo de informação (8,00) foi atingido em 0,455 desvios padrão abaixo da média da habilidade. Este valor corresponde a um coeficiente alfa de Cronbach superior ao valor (constante) obtido da TCT.

Foi comprovada a unidimensionalidade da habilidade, através de uma análise em componentes principais sobre a matriz das correlações tetracóricas dos vários itens, sendo a percentagem da variância total explicada pela primeira componente de 31,6% enquanto que essa percentagem foi de 7,2% para a segunda componente (Figura 53).

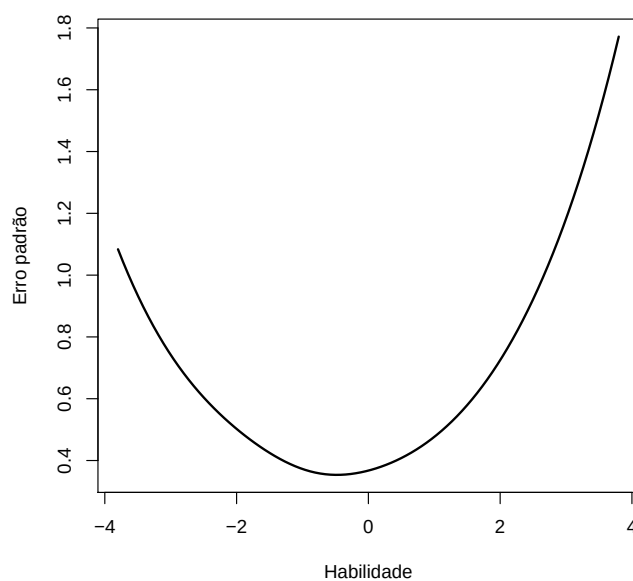


Figura 52: Erro padrão de medição (2013)

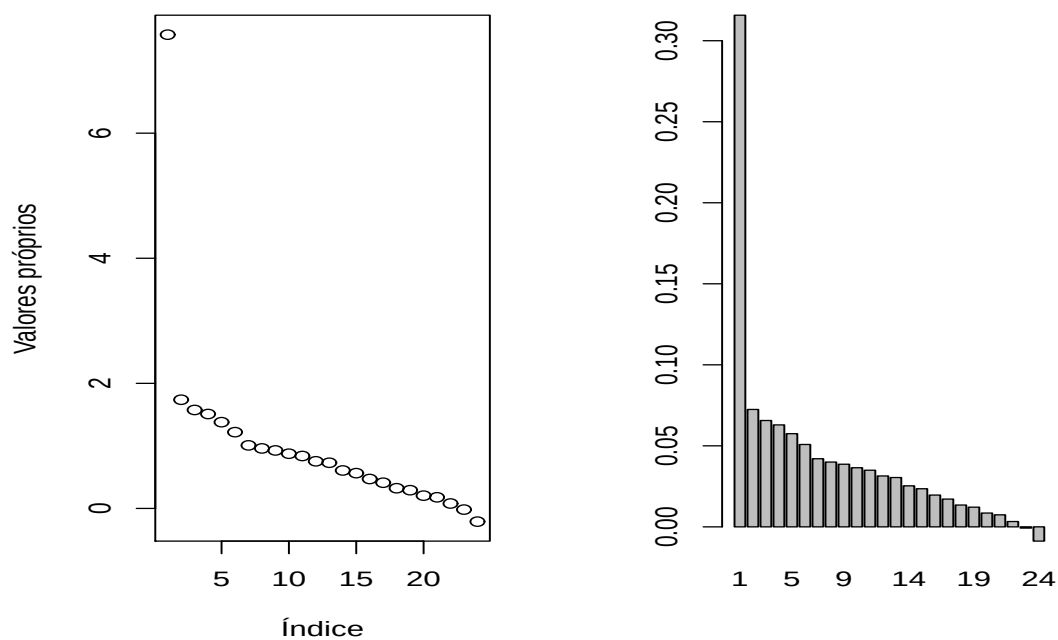


Figura 53: Valores próprios e componentes principais (2013)

A estatística *item-fit* mostrou que apenas os grupos de itens G3 e G16 não se ajustaram corretamente ao modelo seguido. A estatística *person-fit* identificou apenas 2 padrões de resposta como inconsistentes para o modelo estimado.

3.6 Síntese de resultados

O total de perguntas por teste diagnóstico e o número de questões âncora é apresentado na Tabela 30.

Ano	2009	2010	2011	2012	2013
2009	29	28	4	4	4
2010	-	30	4	4	4
2011	-	-	24	24	24
2012	-	-	-	24	24
2013	-	-	-	-	24

Tabela 30: Total de perguntas por teste diagnóstico (diagonal) e número de perguntas âncora

Um item do teste diagnóstico de 2009 foi eliminado por conter duas opções de resposta iguais. Nos anos 2011 a 2013 as 30 questões foram organizadas em 24 grupos de itens.

Na Tabela 31 é apresentada uma análise da qualidade dos itens segundo a Teoria Clássica de Testes.

Ano	N	Parâmetro de dificuldade			Parâmetro de discriminação			Qualidade do teste
		Mediana	P25	P75	Mediana	P25	P75	Alfa de Cronbach
2009	363	47%	39%	57%	0,43	0,35	0,47	0,791
2010	218	46%	37%	55%	0,42	0,37	0,45	0,802
2011	328	50%	46%	60%	0,48	0,44	0,51	0,825
2012	206	61%	49%	70%	0,45	0,36	0,51	0,797
2013	164	57%	48%	70%	0,50	0,46	0,55	0,835

Tabela 31: Análise da qualidade dos itens de acordo com a Teoria Clássica de Testes

A mediana do parâmetro de dificuldade variou entre 46% (2010) e 61% (2012), a mediana do parâmetro de discriminação variou entre 0,42 (2010) e 0,50 (2013) e o coeficiente alfa de Cronbach entre 0,791 (2009) e 0,835 (2013).

Na Tabela 32 é apresentada uma análise da qualidade dos itens segundo o modelo 2PL da Teoria de Resposta ao Item.

Ano	Parâmetro de dificuldade			Parâmetro de discriminação			P(X=1 z=0)		
	Mediana	P25	P75	Mediana	P25	P75	Mediana	P25	P75
2009	0,08	-0,46	0,53	0,90	0,70	1,12	0,47	0,38	0,59
2010	0,22	-0,27	0,88	0,87	0,76	0,96	0,45	0,36	0,57
2011	-0,02	-0,50	0,21	1,03	0,89	1,16	0,51	0,44	0,62
2012	-0,54	-1,22	0,02	0,94	0,74	1,10	0,64	0,49	0,73
2013	-0,33	-0,83	0,08	1,19	1,04	1,32	0,58	0,48	0,74

Tabela 32: Análise da qualidade dos itens de acordo com o modelo 2PL da Teoria de Resposta ao Item

A mediana do parâmetro de dificuldade variou entre -0,54 (2012) e 0,22 (2010), a mediana do parâmetro de discriminação variou entre 0,87 (2010) e 1,19 (2013) e a mediana da probabilidade de acertar o item dada a habilidade igual a zero tomou valores entre 0,45 (2010) e 0,64 (2012).

Na Figura 54 encontra-se a percentagem média de acerto dos estudantes, por ano, em todos os itens do teste e nas questões âncora.

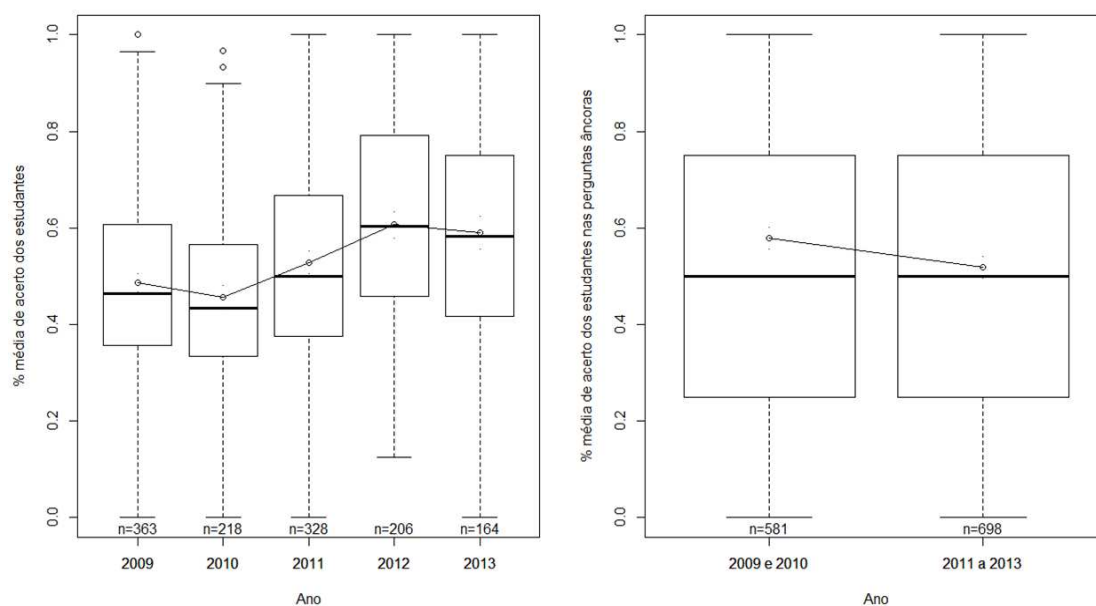


Figura 54: Percentagem média de acerto dos estudantes por ano em todas as perguntas do teste e nas perguntas âncora

Realizou-se uma análise ANOVA, seguida de um teste de comparações múltiplas, tendo-se concluído existirem diferenças estatisticamente significativas entre as habilidades dos diferentes anos. Fazendo uma comparação direta dos resultados obtidos, sem efetuar calibração, existe uma tendência linear positiva dos mesmos. A comparação direta dos resultados obtidos nos diferentes anos sugeriu, portanto, que estes estavam a melhorar de ano para ano.

3.7 Calibração

Uma vez que apenas duas questões, a 6 e a 26, do teste diagnóstico aplicado no ano 2009 foram alteradas, aquando da aplicação do teste diagnóstico do ano 2010, existem 28 questões âncora entre os testes diagnósticos aplicados nestes dois anos.

Relativamente aos testes diagnósticos aplicados no ano 2010 e 2011 existem quatro questões âncora. Considerando o teste diagnóstico aplicado no ano 2010 as questões âncora são a 5, a 7, a 9 e a 14.

Considerando os testes diagnósticos aplicados nos anos 2011, 2012 e 2013, uma vez que os testes foram realizados tendo por base o mesmo banco de questões, consideraram-se as

24 questões como questões âncora. Destas existem 5 questões, a 9, a 11, a 15, a 18 e a 24 que são mesmo fixas e iguais nos testes diagnósticos aplicados nestes três anos.

3.7.1 Calibração usando Teoria Clássica de Testes

A calibração usando Teoria Clássica de Testes foi realizada com recurso ao pacote *equate* do software R.

Na Tabela 33 estão representadas as estatísticas das classificações observadas em cada um dos anos em consideração.

Anos	2009	2010	2011	2012	2013
Média	13,63	13,71	12,68	14,56	14,16
Desvio padrão	5,36	5,54	5,12	4,75	5,17

Tabela 33: Estatísticas das classificações observadas

As distribuições das classificações observadas nos testes e nos itens âncora, para cada um dos anos, encontram-se representadas nas Figuras 55, 56, 57, 58 e 59.

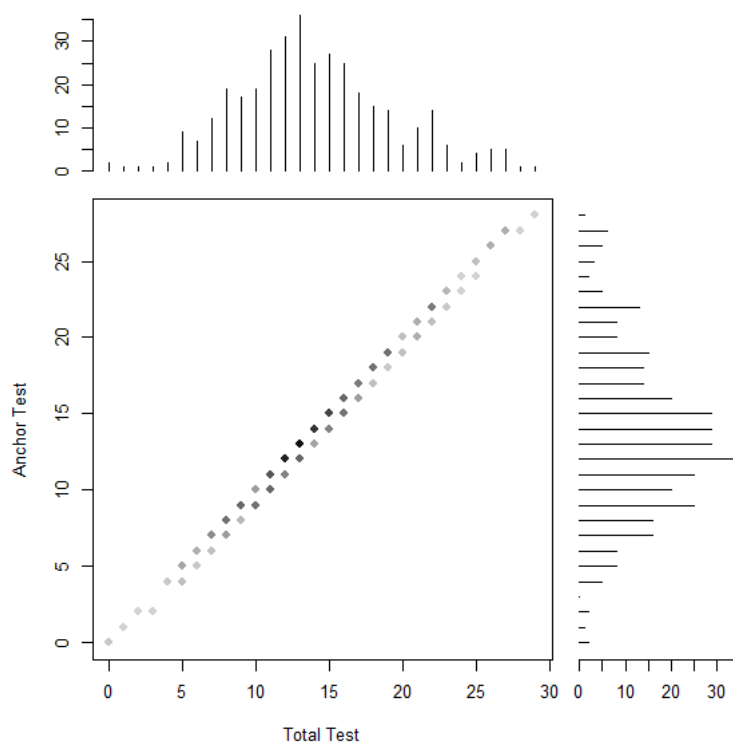


Figura 55: Distribuições das classificações nos testes e nos itens âncora (2009)

Realizou-se o processo de calibração dos resultados tomando como referência o ano 2009. Os resultados dos diferentes anos foram, portanto, calibrados para o ano 2009. Usou-se calibração separada, linear e com grupos não equivalentes. A opção pela calibração linear deveu-se à reduzida dimensão das amostras e ao facto de existirem apenas quatro itens âncora dos testes de 2009 e 2010 para os dois anos seguintes.

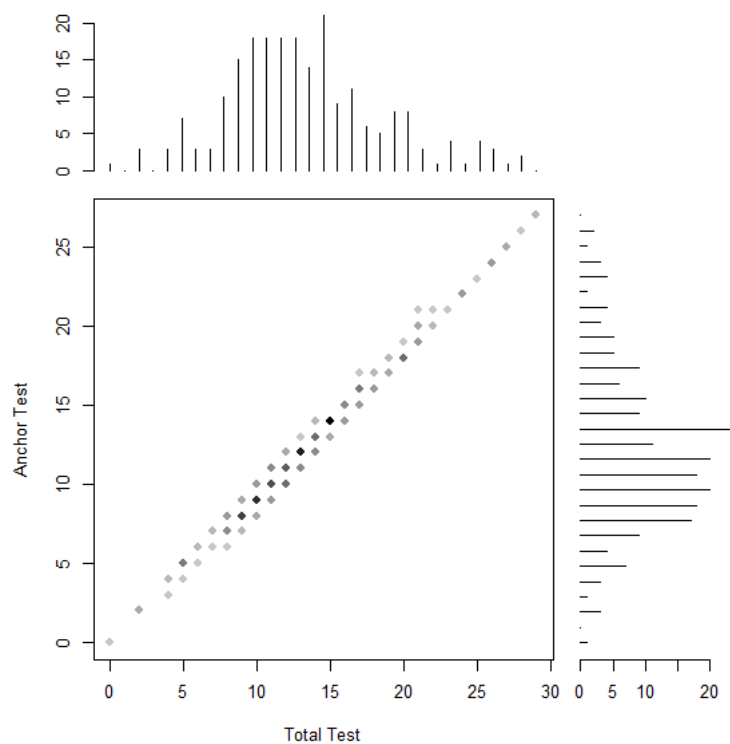


Figura 56: Distribuições das classificações nos testes e nos itens âncora (2010)

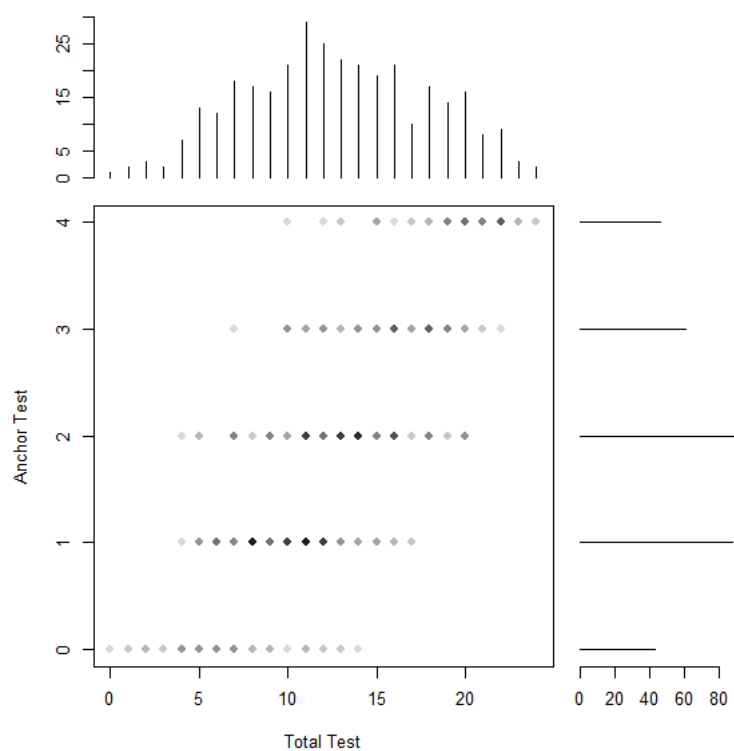


Figura 57: Distribuições das classificações nos testes e nos itens âncora (2011)

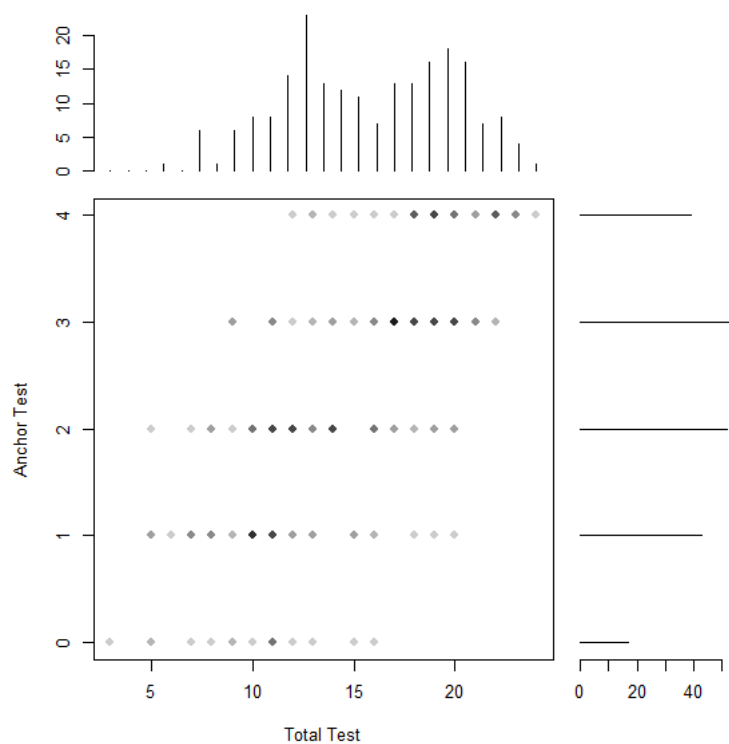


Figura 58: Distribuições das classificações nos testes e nos itens âncora (2012)

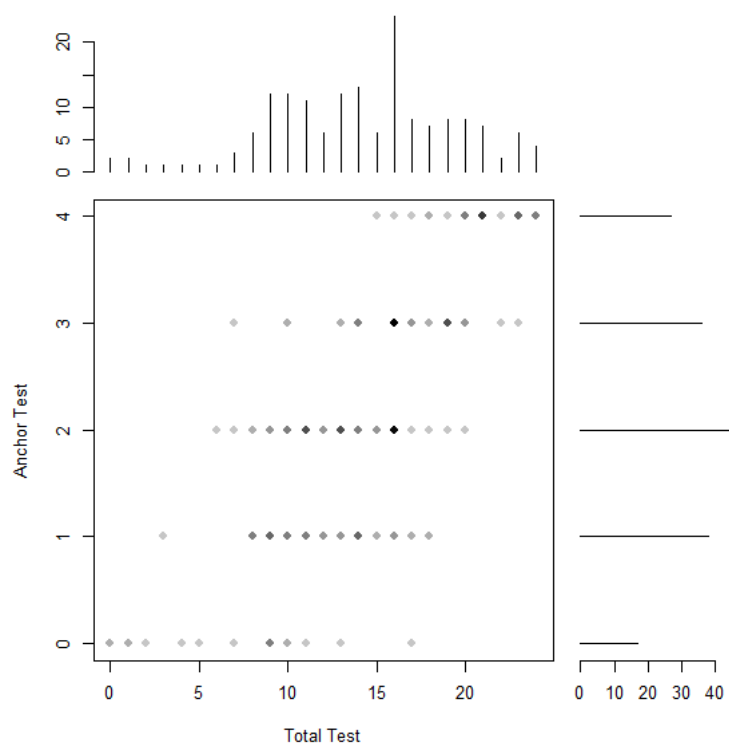


Figura 59: Distribuições das classificações nos testes e nos itens âncora (2013)

A média, o desvio padrão, o máximo e o mínimo das habilidades estimadas para os vários anos, tomando como referência a escala do teste diagnóstico aplicado no ano 2009, encontram-se representados na Tabela 34.

Ano	2009	2010	2011	2012	2013
Média	13,97	12,91	12,52	13,69	13,11
SD	5,32	5,12	5,68	5,60	5,66
Min	0	0,24	-1,55	0,05	-2,38
Max	29	27,04	25,07	24,83	23,87

Tabela 34: Estatísticas das habilidades estimadas usando TCT, referência 2009

Realizou-se uma análise ANOVA, seguida de um teste de comparações múltiplas, tendo-se concluído existirem diferenças estatisticamente significativas entre as habilidades dos diferentes anos. Contudo não existe uma tendência temporal estatisticamente significativa. Em média o grupo de estudantes que revelou maior habilidade foi o que realizou o teste diagnóstico aplicado no ano 2009, seguindo-se os que realizaram o teste diagnóstico aplicado no ano 2012 e os dos anos 2013 e 2010 com uma habilidade média muito próxima. O grupo de estudantes com menor habilidade média foi o que realizou o teste diagnóstico no ano 2011. Na Figura 60 é apresentada a evolução temporal da habilidade média.

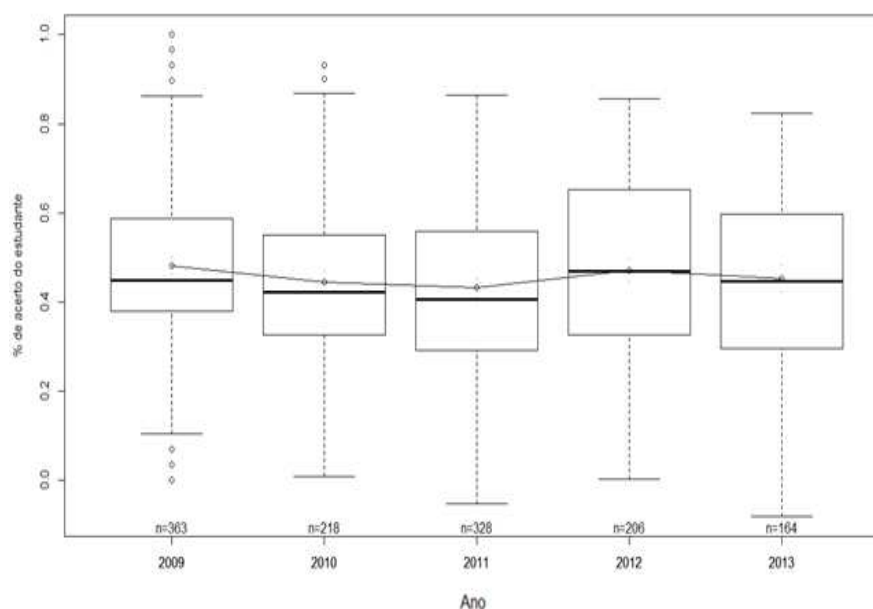


Figura 60: Evolução temporal da habilidade média estimada usando TCT

Relativamente ao desvio padrão este é praticamente constante ao longo dos anos em análise.

Na Tabela 35 são apresentadas classificações calibradas para os vários anos tendo como referência o teste diagnóstico realizado em 2009.

Escala de Referência	2010 → 2009	2011 → 2009	2012 → 2009	2013 → 2009
0	0,24	-1,55	-3,49	-2,38
1	1,16	-0,44	-2,31	-1,29
2	2,09	0,67	-1,13	-0,19
3	3,01	1,78	0,05	0,90
4	3,94	2,89	1,23	2,00
5	4,86	4,00	2,41	3,09
6	5,79	5,10	3,59	4,18
7	6,71	6,21	4,77	5,28
8	7,63	7,32	5,95	6,37
9	8,56	8,43	7,13	7,47
10	9,48	9,54	8,31	8,56
11	10,41	10,65	9,49	9,65
12	11,33	11,76	10,67	10,75
13	12,26	12,87	11,85	11,84
14	13,18	13,98	13,03	12,93
15	14,10	15,08	14,21	14,03
16	15,03	16,19	15,39	15,12
17	15,95	17,30	16,57	16,22
18	16,88	18,41	17,75	17,31
19	17,80	19,52	18,93	18,40
20	18,72	20,63	20,11	19,50
21	19,65	21,74	21,29	20,59
22	20,57	22,85	22,47	21,69
23	21,50	23,96	23,65	22,78
24	22,42	25,07	24,83	23,87
25	23,35	-	-	-
26	24,27	-	-	-
27	25,19	-	-	-
28	26,12	-	-	-
29	27,04	-	-	-
30	27,97	-	-	-

Tabela 35: Classificações calibradas tendo como referência o ano 2009

3.7.2 Calibração usando Teoria de Resposta ao Item

A calibração usando Teoria de Resposta ao Item foi realizada com recurso ao pacote plink do software R.

Os resultados dos diferentes anos foram calibrados para o ano adjacente de 2013 para 2009. O método usado no processo de calibração foi o média/média.

Os valores das constantes de ligação obtidos, tomando como escala de referência a do teste diagnóstico aplicado no ano 2009, para o método média/média encontram-se representados na Tabela 36.

Anos	2010/2009		2011/2010		2012/2011		2013/2012	
Constantes	A	B	A	B	A	B	A	B
Média/Média	0,917	-0,261	1,081	-0,392	0,887	0,500	1,273	-0,203

Tabela 36: Constantes de ligação referência 2009

Na Tabela 37 indicam-se as constantes de calibração obtidas com a Teoria Clássica de Testes e com a Teoria de Resposta ao Item.

Teoria Clássica de Testes			Teoria de Resposta ao Item		
Conversão	A	B	Conversão	A	B
2010 → 2009	0,924	0,240	2010 → 2009	0,917	-0,261
2011 → 2009	1,109	-1,555	2011 → 2010	1,081	-0,392
2012 → 2009	1,180	-3,489	2012 → 2011	0,887	0,500
2013 → 2009	1,094	-2,340	2013 → 2012	1,273	-0,203

Tabela 37: Constantes de calibração de acordo com a TCT e a TRI

A média, o desvio padrão, o máximo e o mínimo das habilidades estimadas para os vários anos, tomando como referência a escala do teste diagnóstico aplicado no ano 2009, encontram-se representados na Tabela 38. Considerou-se para os diversos anos uma escala para a habilidade de -4 a 4 logits e o método média/média.

Anos	2009	2010	2011	2012	2013
Média	0	-0,261	-0,620	-0,125	-0,304
Desvio padrão	2,739	2,510	2,713	2,406	3,064
Mínimo	-4	-3,927	-4,583	-3,640	-4,779
Máximo	4	3,406	3,343	3,390	4,172

Tabela 38: Estatísticas das habilidades estimadas usando TRI, referência 2009

Realizou-se uma análise ANOVA, seguida de um teste de comparações múltiplas, tendo-se concluído existirem diferenças estatisticamente significativas entre as habilidades dos diferentes anos. Existe também uma tendência linear negativa estatisticamente significativa. Em média o grupo de estudantes que revelou maior habilidade foi o que realizou o teste diagnóstico aplicado no ano 2009. O grupo de estudantes com menor habilidade média foi

o que realizou o teste diagnóstico no ano 2011. Na Figura 61 é apresentada a evolução temporal da habilidade média.

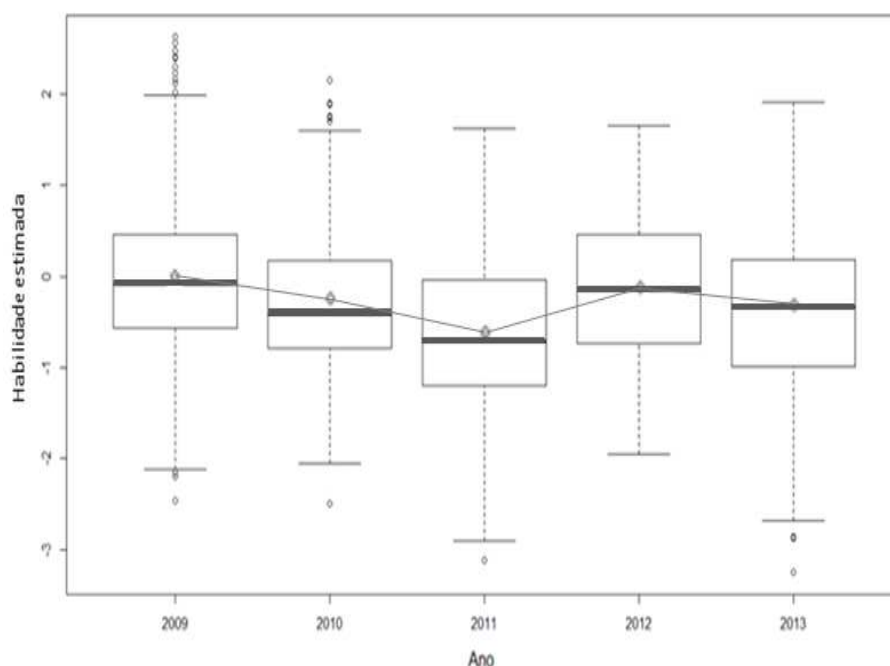


Figura 61: Evolução temporal da habilidade média estimada usando TRI

Os resultados obtidos realizando calibração usando Teoria de Resposta ao Item são muito semelhantes aos que se obtiveram no processo de calibração usando Teoria Clássica de Testes. Quando se ordena os diferentes grupos de estudantes, de acordo com a sua habilidade média estimada, a diferença ocorre entre os anos 2010 e 2013, contudo os valores das habilidades médias destes anos são bastante próximos Figura 62.

Com a Teoria Clássica de Testes não há evidência estatisticamente significativa de uma tendência linear dos resultados, usando Teoria de Resposta ao Item há evidência estatisticamente significativa de uma tendência linear negativa.

Relativamente ao desvio padrão este é praticamente constante ao longo dos anos em análise, tal como havia sucedido na calibração usando Teoria Clássica de Testes.

Na Tabela 39 encontra-se representada a classificação verdadeira dado um determinado valor da habilidade θ .

Na Teoria de Resposta ao Item a classificação verdadeira, dado um determinado valor da habilidade, corresponde à soma das probabilidades esperadas dos vários itens para o valor da habilidade em causa.

Os valores indicados na Tabela 39 resultam de efetuar calibração usando a classificação verdadeira, ou seja, de determinar o valor da habilidade associado a uma determinada classificação verdadeira na escala de referência e, posteriormente, determinar as correspondentes classificações verdadeiras nos testes diagnósticos aplicados nos restantes anos.

Para mais fácil interpretação, apresenta-se na Tabela 40 a classificação verdadeira na escala 0-20 dado um determinado valor da habilidade θ .

θ	2009	2010	2011	2012	2013
-4,16	1	0,90	1,13	1,47	1,10
-3,07	2	1,93	2,71	2,79	2,72
-2,49	3	3,00	4,16	4,00	4,20
-2,10	4	4,10	5,49	5,16	5,55
-1,79	5	5,22	6,72	6,27	6,81
-1,53	6	6,36	7,89	7,35	8,00
-1,30	7	7,50	9,02	8,41	9,14
-1,10	8	8,64	10,11	9,45	10,23
-0,91	9	9,77	11,16	10,47	11,27
-0,73	10	10,89	12,18	11,48	12,27
-0,55	11	11,99	13,16	12,47	13,23
-0,39	12	13,08	14,11	13,44	14,15
-0,22	13	14,15	15,01	14,39	15,03
-0,05	14	15,20	15,88	15,31	15,88
0,12	15	16,23	16,70	16,20	16,69
0,29	16	17,26	17,48	17,07	17,47
0,47	17	18,27	18,23	17,89	18,22
0,65	18	19,27	18,94	18,69	18,93
0,84	19	20,26	19,62	19,44	19,61
1,05	20	21,24	20,26	20,15	20,25
1,28	21	22,20	20,86	20,81	20,86
1,52	22	23,16	21,42	21,42	21,42
1,80	23	24,10	21,95	21,98	21,94
2,12	24	25,03	22,43	22,49	22,42
2,50	25	25,95	22,87	22,93	22,86
2,99	26	26,87	23,26	23,31	23,25
3,69	27	27,80	23,60	23,64	23,58
5,03	28	28,82	23,87	23,89	23,86

Tabela 39: Classificação verdadeira dado θ

θ	2009	2010	2011	2012	2013
-4,16	0,69	0,60	0,94	1,23	0,92
-3,07	1,38	1,29	2,26	2,32	2,27
-2,49	2,07	2,00	3,47	3,34	3,50
-2,10	2,76	2,73	4,57	4,30	4,62
-1,79	3,45	3,48	5,60	5,22	5,67
-1,53	4,14	4,24	6,58	6,12	6,67
-1,30	4,83	5,00	7,51	7,01	7,62
-1,10	5,52	5,76	8,42	7,87	8,52
-0,91	6,21	6,52	9,30	8,73	9,39
-0,73	6,90	7,26	10,15	9,57	10,23
-0,55	7,59	8,00	10,97	10,39	11,03
-0,39	8,28	8,72	11,76	11,20	11,79
-0,22	8,97	9,43	12,51	11,99	12,53
-0,05	9,66	10,13	13,23	12,76	13,23
0,12	10,34	10,82	13,92	13,50	13,91
0,29	11,03	11,50	14,57	14,22	14,56
0,47	11,72	12,18	15,19	14,91	15,18
0,65	12,41	12,85	15,78	15,57	15,78
0,84	13,10	13,51	16,35	16,20	16,34
1,05	13,79	14,16	16,88	16,79	16,88
1,28	14,48	14,80	17,38	17,34	17,38
1,52	15,17	15,44	17,85	17,85	17,85
1,80	15,86	16,07	18,29	18,32	18,29
2,12	16,55	16,69	18,69	18,74	18,69
2,50	17,24	17,30	19,06	19,11	19,05
2,99	17,93	17,91	19,38	19,43	19,37
3,69	18,62	18,53	19,67	19,70	19,65
5,03	19,31	19,21	19,89	19,90	19,88

Tabela 40: Classificação verdadeira (escala 0-20) dado θ

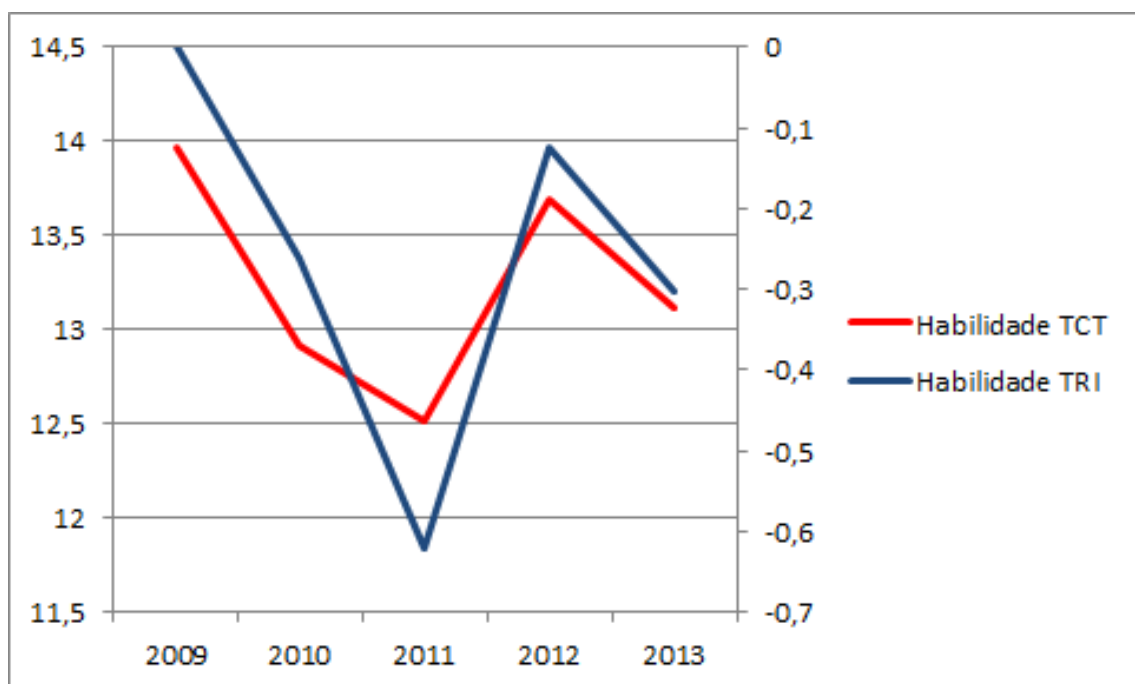


Figura 62: Comparação da evolução temporal da habilidade média estimada usando TCT e TRI

Capítulo 4

Avaliação do impacto do treino

4.1 Descrição do problema

Pretendeu-se avaliar o impacto do treino realizado pelos alunos na plataforma online nos resultados obtidos num teste realizado no âmbito da avaliação da unidade curricular de Cálculo. Para tal recorreu-se a um modelo de regressão linear múltipla.

4.2 Descrição dos dados

Foi recolhida informação referente a 108 estudantes e foram registadas 4 variáveis, a saber:

Curso: Curso frequentado (CC - Ciência de Computadores; CE - Ciências de Engenharia; MIERSI - Mestrado Integrado em Engenharia de Redes e Sistemas Informáticos);

TD: Classificação obtida no teste diagnóstico (escala 0-30);

Tempo: Tempo despendido na plataforma (em minutos);

Avaliacao: Classificação obtida no teste de Cálculo (escala 0-30).

4.3 Descrição estatística dos dados

4.3.1 Variável dependente: Avaliacao

A média da variável dependente, Avaliacao, foi de 22,17, a mediana de 23,33 e os quartis 18,92 e 26. O mínimo foi 4,67 e o máximo 30.

Na Figura 63 são apresentadas representações gráficas da variável dependente.

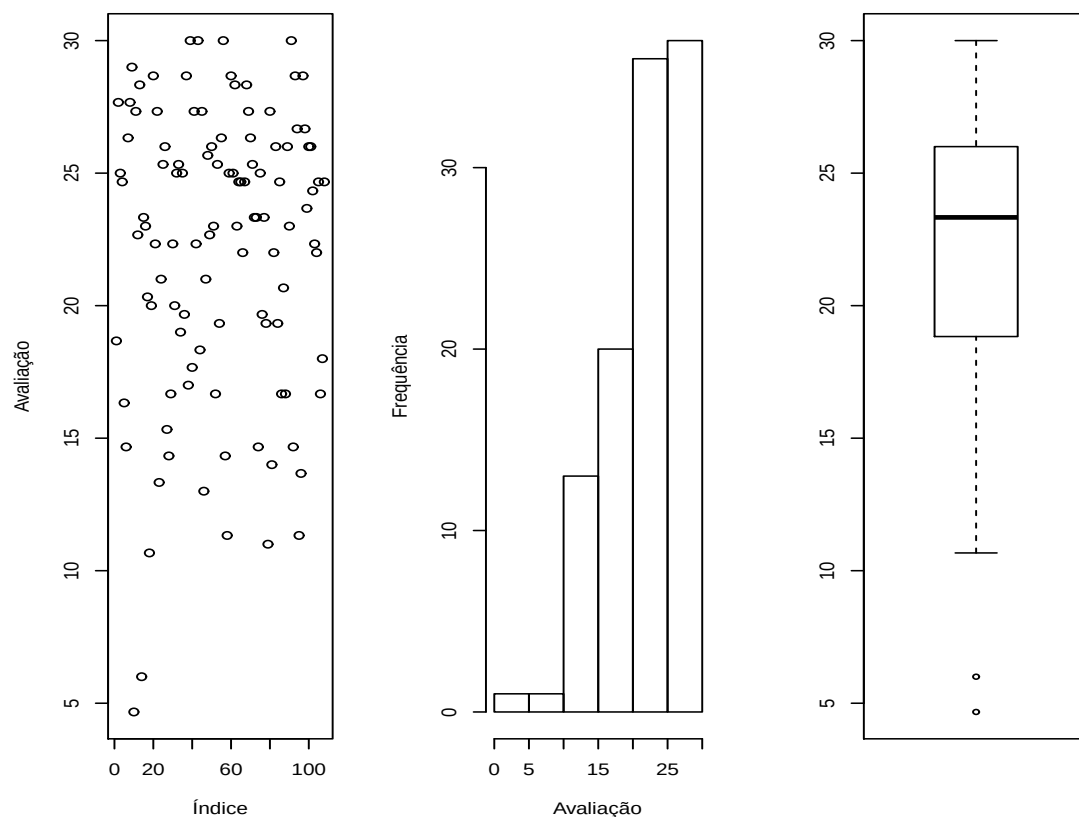


Figura 63: Representações gráficas da variável dependente: Avaliacao

4.3.2 Variáveis explicativas

4.3.2.1 Curso

No que diz respeito à variável Curso registaram-se 32 estudantes (30%) do curso de Ciência de Computadores, 14 (13%) de Ciências de Engenharia e 62 (57%) do Mestrado Integrado em Engenharia de Redes e Sistemas Informáticos. Na Tabela 41 apresentam-se estatísticas numéricas da variável dependente, Avaliacao, em função do curso frequentado pelos estudantes.

Curso	Mín	p_{25}	p_{50}	Média	p_{75}	Máx
Ciência de Computadores	10,67	17,67	23	21,54	25,75	28,67
Ciências de Engenharia	11	23	25,5	23,88	27,16	30
Mest. Int. Eng. Redes e Sist. Informáticos	4,67	19,08	23	22,11	26	30

Tabela 41: Estatísticas numéricas descritivas da variável Avaliacao em função da variável Curso

Na Figura 64 são apresentadas representações gráficas da variável explicativa Curso.

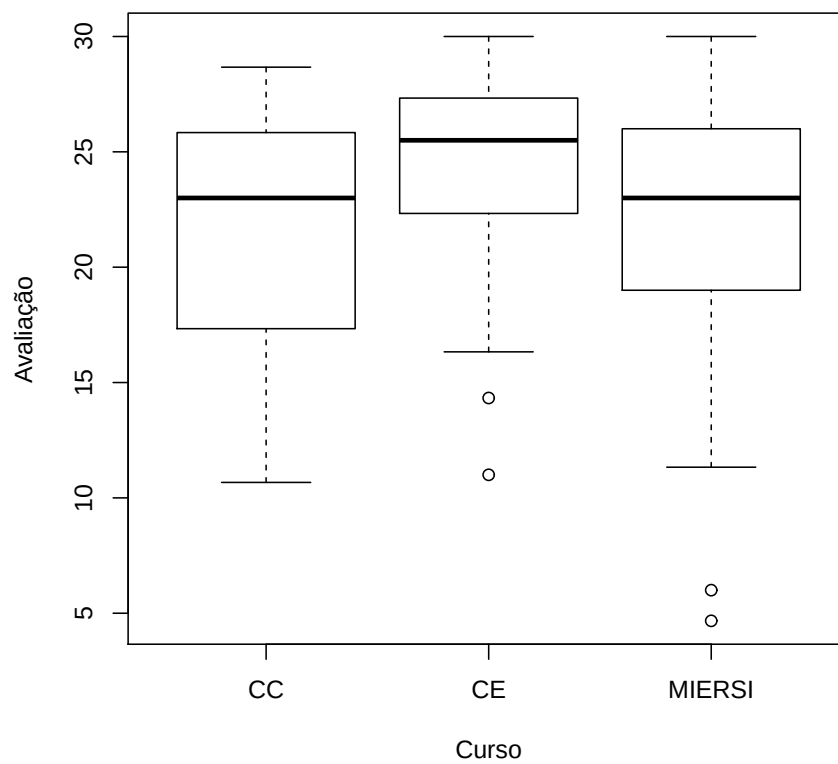


Figura 64: Representações gráficas da variável explicativa Curso

4.3.2.2 TD

A média da variável TD foi de 15,99, a mediana de 16,33 e os quartis 11,67 e 20,67. O mínimo foi 0 e o máximo 28,67.

Na Figura 65 são apresentadas representações gráficas da variável explicativa TD.

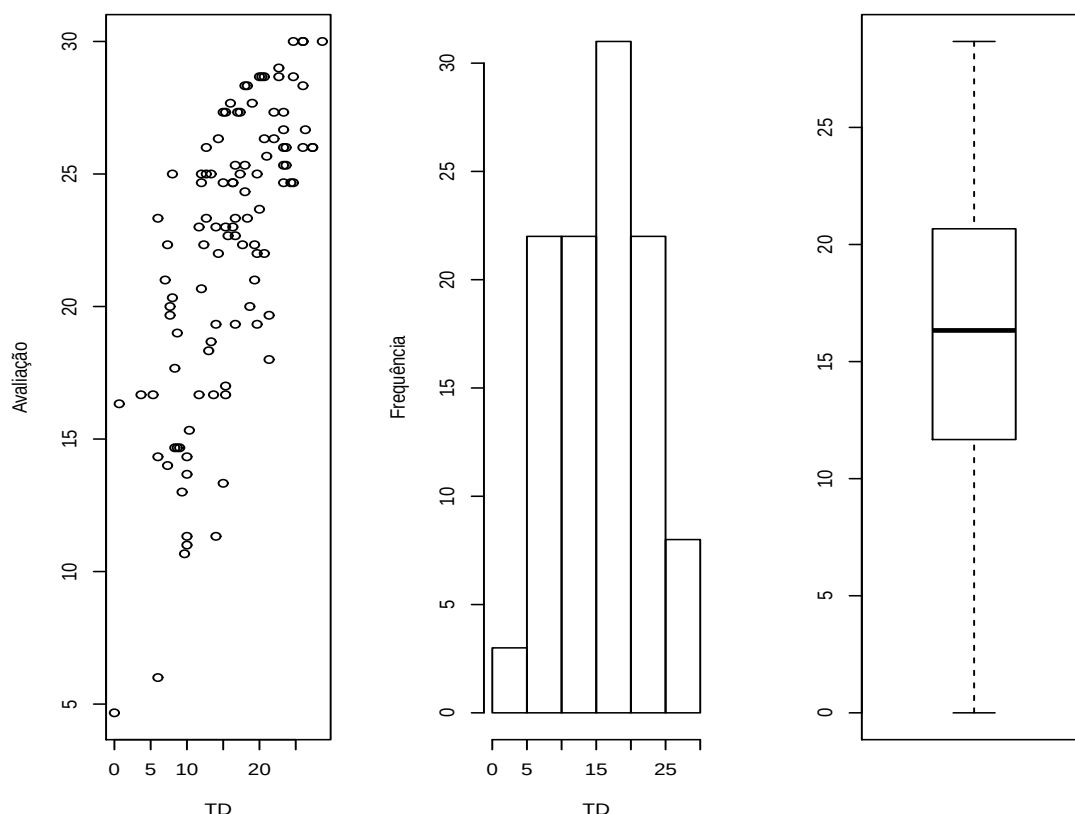


Figura 65: Representações gráficas da variável explicativa TD

4.3.2.3 Tempo

A média da variável Tempo foi de 189,7 minutos, a mediana de 129,4 e os quartis 64,08 e 273,3. O mínimo foi 0 e o máximo 1174.

De modo a estudar o impacto do treino na plataforma na avaliação da unidade curricular de Cálculo a variável Tempo foi dicotomizada considerando-se as seguintes categorias: 0 (Treino insuficiente) e 1 (Treino suficiente). A distinção das categorias foi realizada considerando que os estudantes que estiveram menos de 120 minutos na plataforma não treinaram o suficiente. Designou-se a nova variável por Treino. Realizada esta dicotomização registaram-se 50 alunos que não treinaram o suficiente e 58 que treinaram o suficiente. Na Tabela 42 apresentam-se estatísticas numéricas da variável Avaliação em função da variável Treino. Na Figura 66 são apresentadas representações gráficas da variável Avaliação em função da variável Treino.

Tempo	Mín	p_{25}	p_{50}	Média	p_{75}	Máx
0	4,67	16,67	20,84	20,68	25,33	30
1	6	22	24,84	23,45	26,58	30

Tabela 42: Estatísticas numéricas da variável Avaliacao em função da variável Treino

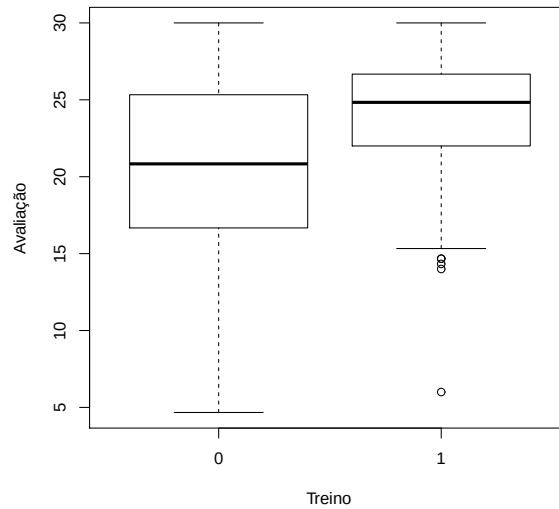


Figura 66: Representações gráficas da variável Avaliacao em função da variável Treino

4.4 Ajustamento do modelo

Considerando o objetivo do estudo e não tendo nenhuma indicação extra sobre as variáveis em causa, começámos por analisar o modelo de regressão linear completo, ou seja, considerando todas as variáveis explicativas recolhidas.

Tal modelo, que designámos por modelo 1, apresenta a seguinte equação:

$$\hat{y} = \beta_0 + \beta_1 TD + \beta_2 Treino + \beta_3 CE + \beta_4 MIERSI + e$$

em que TD representa a variável TD (classificação obtida no teste diagnóstico), Treino representa a variável indicatriz da existência de treino (isto é, Treino = 0 se o treino foi insuficiente e Treino = 1 se o treino foi suficiente), CE representa a variável indicatriz do Curso de Ciências de Engenharia (isto é, CE = 0 se o curso frequentado pelo aluno não é Ciências de Engenharia e CE = 1 se o curso frequentado pelo aluno é Ciências de Engenharia) e MIERSI representa a variável indicatriz do Curso de Mestrado Integrado em Engenharia de Redes e Sistemas Informáticos (isto é, MIERSI = 0 se o curso frequentado pelo aluno não é Mestrado Integrado em Engenharia de Redes e Sistemas Informáticos e MIERSI = 1 se o curso frequentado pelo aluno é Mestrado Integrado em Engenharia de Redes e Sistemas Informáticos).

Sob a condição $e \sim N(0, \sigma^2 Id)$ de os erros serem independentes e normalmente identicamente distribuídos.

O modelo 1 possui um coeficiente de determinação de 0,561, um coeficiente de determinação ajustado de 0,544, o valor da estatística F, que compara o modelo nulo com o modelo atual, foi de 32,86 e o valor-p inferior a 0,001.

Da análise do modelo obtido (Tabela 43) e da tabela ANOVA do mesmo (Tabela 44) foi possível concluir que a variável explicativa Curso não era estatisticamente significativa para o modelo.

Coeficientes	Estimativa	Erro St.	Valor t	P(> t)
Constante	10,442	1,182	8,834	<0,001
CursoCE	1,555	1,222	1,273	0,206
CursoMIERSI	1,044	0,809	1,291	0,199
TD	0,605	0,056	10,732	<0,001
Treino	2,343	0,739	3,172	<0,001

Tabela 43: Modelo 1

	G. L.	Soma de Quad.	Média de Quad.	Valor F	P(>F)
Curso	2	53,84	26,92	1,96	0,146
TD	1	1612,15	1612,15	117,44	<0,001
Treino	1	138,14	138,14	10,06	0,002
Resíduos	103	1413,97	13,73		

Tabela 44: ANOVA Modelo 1

Da análise da Figura 67 é possível concluir que não existem *outliers*. Um *outlier* é uma observação que não se ajusta ao modelo encontrado apresentando resíduos grandes quando comparados com os resíduos das restantes observações. Geralmente, assume-se que as observações cujos resíduos estandardizados são superiores a 3,3 são *outliers*.

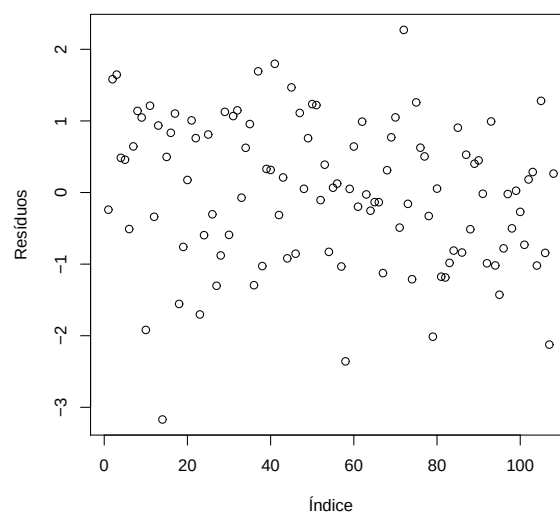


Figura 67: Resíduos (Modelo 1)

A multicolinearidade pode ser averiguada através da regressão de cada uma das variáveis explicativas sobre as restantes. Para tal considerámos o VIF (fator de inflação da variância) associado a cada uma das variáveis explicativas. Não existe multicolinearidade elevada, uma vez que nenhuma das variáveis explicativas apresentou um valor de VIF superior a 10 (Tabela 45).

CursoCE	CursoMIERSI	TD	Treino >2h
1,325	1,257	1,005	1,067

Tabela 45: VIF Modelo 1

Considerando o modelo obtido eliminando a variável explicativa Curso, por não ser estatisticamente significativa, o qual designámos por modelo 2, obtiveram-se os valores indicados na Tabela 46.

Coefficientes	Estimativa	Erro St.	Valor t	P(> t)
Constante	11,216	1,031	10,882	<0,001
TD	0,601	0,056	10,667	<0,001
Treino	2,515	0,716	3,511	<0,001

Tabela 46: Modelo 2

O modelo 2 possui um coeficiente de determinação de 0,551, um coeficiente de determinação ajustado de 0,542, estatística F 64,4 e valor-p inferior a 0,001.

	G. L.	Soma de Quad.	Média de Quad.	Valor F	P(>F)
TD	1	1603,14	1603,14	116,47	<0,001
Treino	1	169,68	169,68	12,33	<0,001
Resíduos	105	1445,27	13,76		

Tabela 47: ANOVA Modelo 2

Da análise do modelo anterior (Tabela 46) e da tabela ANOVA do mesmo (Tabela 47) foi possível concluir que as variáveis explicativas TD e Treino eram estatisticamente significativas para o modelo.

O modelo 2 possui um coeficiente de determinação ajustado razoável pelo que revelou um ajustamento satisfatório aos dados.

Seja y a variável Avaliacao, TD a variável TD (classificação obtida no teste diagnóstico), Treino a variável indicatriz da existência de treino (isto é, Treino = 0 se o treino foi insuficiente e Treino = 1 se o treino foi suficiente). A equação matemática do modelo 2 é:

$$\hat{y} = 11,22 + 0,60 TD + 2,52 Treino.$$

4.5 Diagnóstico e estudo do modelo final

4.5.1 *Outliers*

Da análise da Figura 68 é possível concluir que não existem *outliers*, uma vez que não se identificaram observações cujos resíduos standardizados fossem superiores a 3,3.

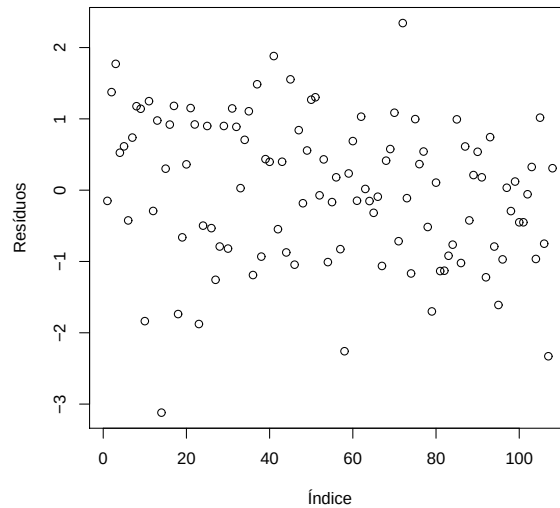


Figura 68: Resíduos (Modelo 2)

4.5.2 *Leverages*

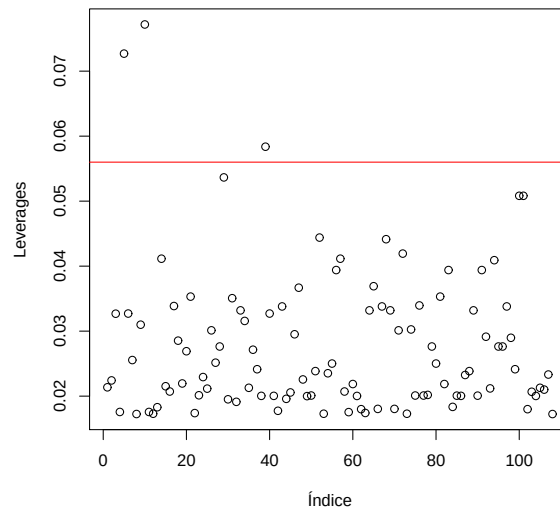
Na regressão múltipla podem existir observações que não sendo *outliers* influenciam demasiado a estimação dos parâmetros do modelo. Diz-se que estas observações têm *leverages* altas. Resumidamente, a *leverage* de uma observação reflete a possibilidade de essa observação ser um *outlier* para as variáveis explicativas e/ou influenciar o seu próprio valor ajustado, sendo necessário analisar a influência exercida por estas observações na estimação dos coeficientes do modelo.

Diz-se que uma observação tem *leverage* alta se a sua *leverage* é superior a $\frac{2(p+1)}{n}$, sendo p o número de variáveis explicativas do modelo e n o número total de observações. No caso do modelo 2 existem três observações com *leverage* alta (Figura 69).

Para análise destas observações é necessário determinar a sua influência.

4.5.3 Observações influentes

Uma observação diz-se influente se tem uma grande influência sobre os parâmetros de regressão estimados, isto é, se o valor de pelo menos um desses parâmetros determinado a partir do conjunto de dados incluindo e excluindo a observação influente é muito diferente. Uma observação influente pode ou não ser um *outlier* e pode ou não ter uma *leverage* grande, mas tenderá a ter pelo menos uma destas propriedades. A estatística habitualmente usada para avaliar a influência de uma observação é a distância de Cook. Considera-se que os

Figura 69: *Leverages* (Modelo 2)

pontos com distância de Cook superior a 1 podem ser pontos influentes pelo que o seu impacto no ajustamento do modelo deve ser averiguado.

Como pode concluir-se da análise da Figura 70, nenhuma observação apresenta distância de Cook superior a 1.

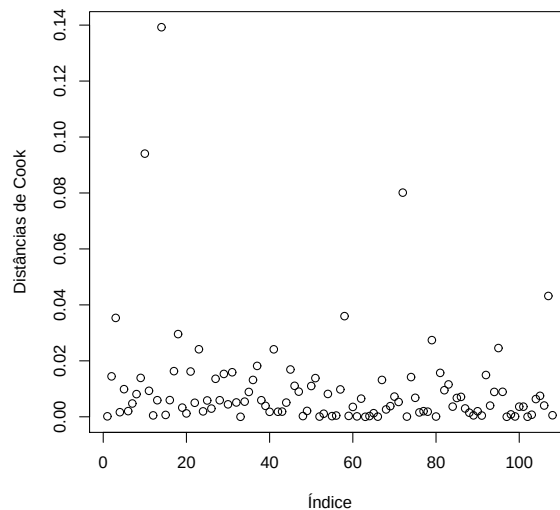
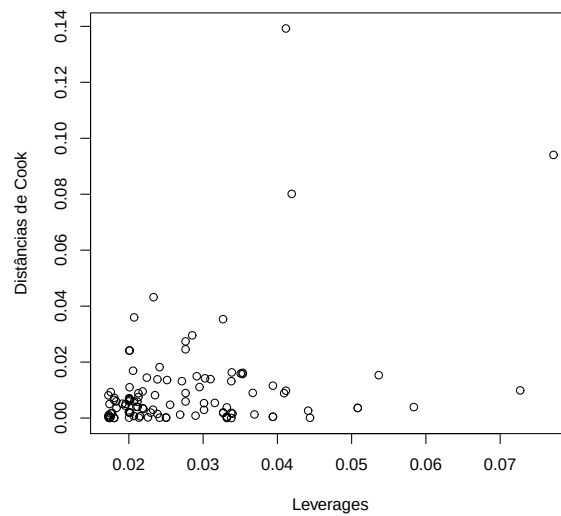


Figura 70: Distâncias de Cook (Modelo 2)

Para análise das *leverages* altas que apresentem distância de Cook elevada considere-se a Figura 71. Como nenhuma das observações possui valores elevados em ambos os eixos, não existem observações problemáticas.

Figura 71: *Leverages* / Distâncias de Cook (Modelo 2)

Ainda em relação à detecção de *outliers* é frequente recorrer-se ao gráfico seminormal (Figura 72).

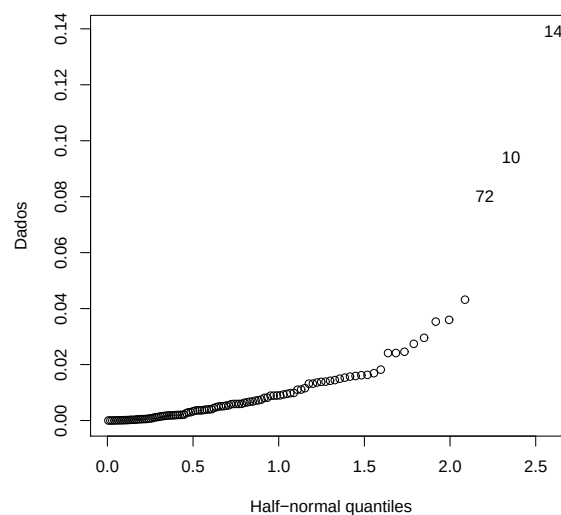


Figura 72: Seminormal (Modelo 2)

Tal como já havia sido evidenciado nos resíduos estandardizados não existem *outliers*.

4.5.4 Homocedasticidade

Uma das condições subjacente a um modelo de regressão linear é a de que os resíduos devem apresentar variância constante. A esta condição chama-se homocedasticidade. Caso

tal não suceda diz-se que estamos na presença de heterocedasticidade.

A verificação da homocedasticidade pode ser feita pela análise gráfica dos resíduos estandardizados contra os valores ajustados (Figura 73).

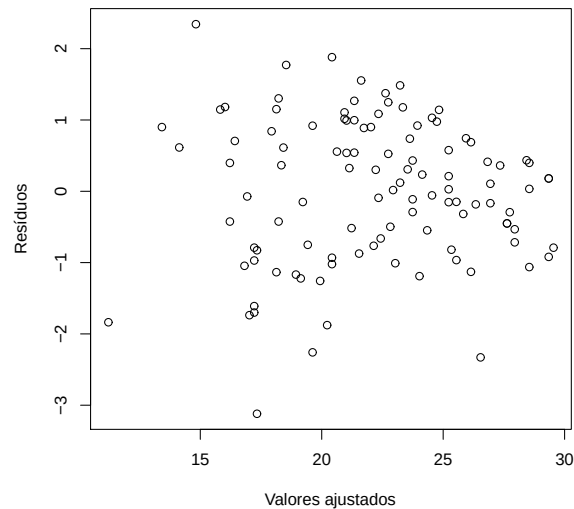


Figura 73: Valores ajustados / Resíduos estandardizados (Modelo 2)

Como os pontos parecem estar dispostos aleatoriamente, ou seja, sem nenhum padrão definido e de forma simétrica em torno do 0, não há evidência de heterocedasticidade.

Uma outra forma de analisar a homocedasticidade é através do gráfico dos valores ajustados contra a resposta (Figura 74).

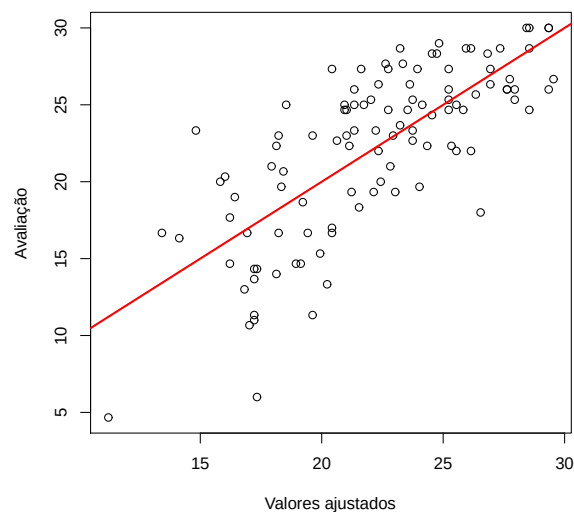


Figura 74: Valores ajustados / Resposta (Modelo 2)

Como as observações se distribuem pela (imaginada) reta de regressão de igual forma ao longo dos valores da resposta não há evidência de heterocedasticidade.

4.5.5 Normalidade dos resíduos estandardizados

Para análise da normalidade dos resíduos considerem-se as representações gráficas presentes na Figura 75. Estas representações indiciam a existência de normalidade dos resíduos estandardizados.

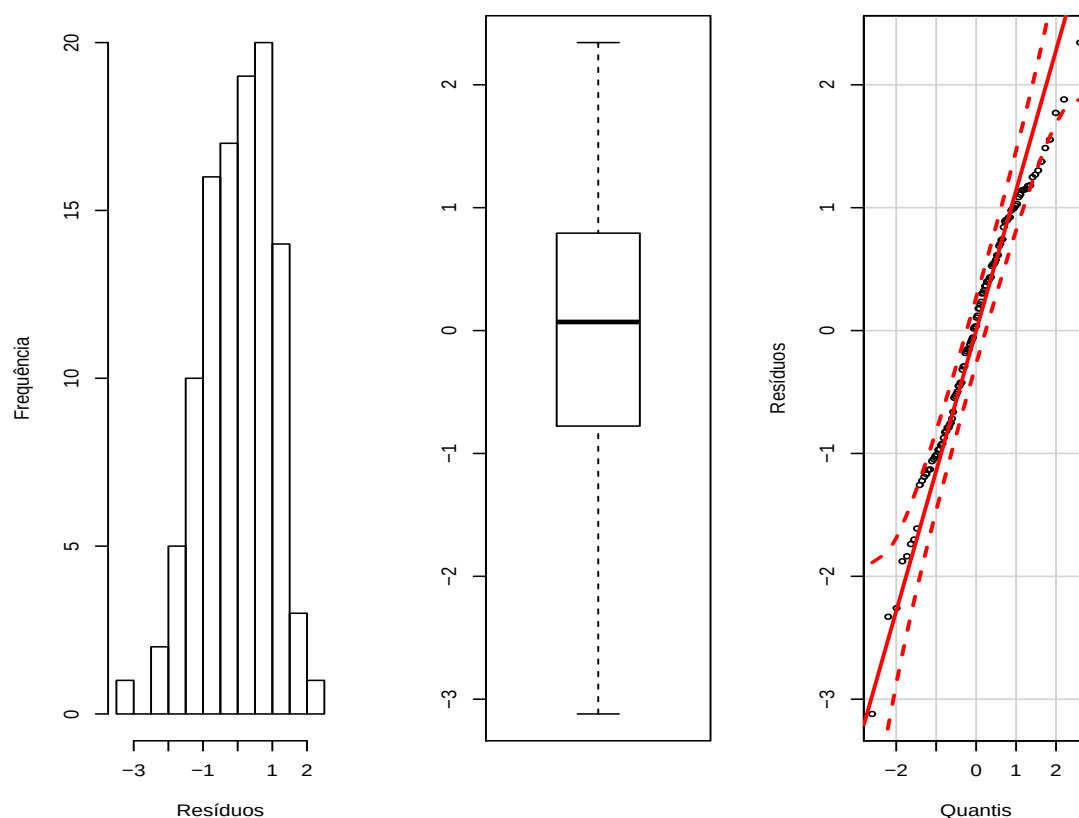


Figura 75: Análise da normalidade dos resíduos estandardizados (Modelo 2)

4.6 Interações

Caso se pretenda melhorar a qualidade do ajustamento de um modelo aos dados pode ter-se em consideração a inclusão no modelo de interações entre as variáveis explicativas.

Nas Figuras 76 e 77 são apresentadas graficamente as interações entre a variável TD e as variáveis Curso e Treino, respetivamente.

Confirmou-se, numericamente, que nenhuma das interações entre as variáveis explicativas anteriormente referidas se revelou estatisticamente significativa.

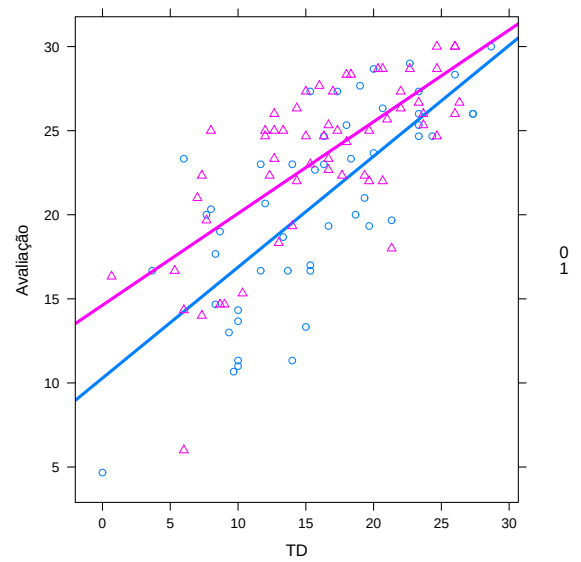


Figura 76: Interação entre as variáveis TD e Treino

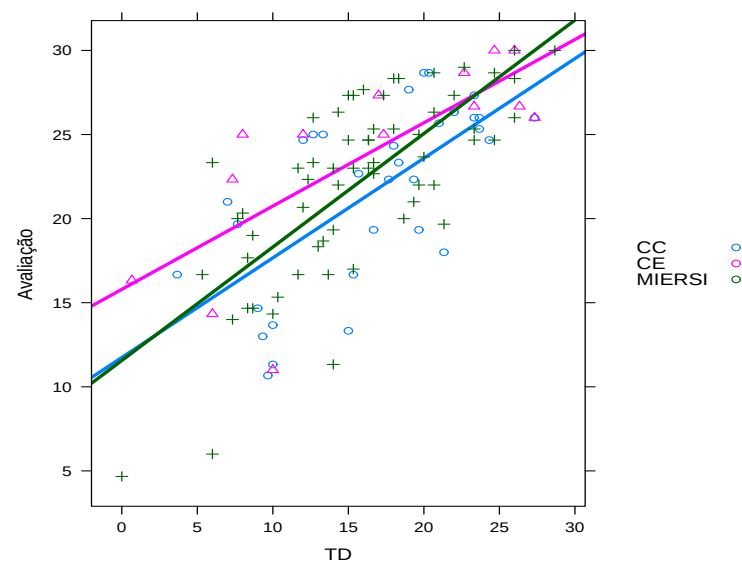


Figura 77: Interação entre as variáveis TD e Curso

Capítulo 5

Conclusões

Com a realização deste trabalho foi possível concluir que os testes diagnósticos de Matemática aplicados pela Faculdade de Ciências da Universidade do Porto, essencialmente a alunos do primeiro ano das suas licenciaturas, de 2009 a 2013, apesar de incluírem itens que aferem diferentes conteúdos matemáticos, medem um construto geral, já que foi comprovada a unidimensionalidade da habilidade.

Uma vez que o modelo logístico com dois parâmetros (2PL) da Teoria de Resposta ao Item foi o que revelou melhor ajustamento aos dados não se pode assumir que todas as questões discriminam de igual forma. Por esta razão os itens devem ser ponderados de acordo com a sua capacidade de discriminação, não devendo ser atribuída a mesma cotação a todos os itens como tem sido feito.

De um modo geral, os itens presentes nos testes diagnósticos foram bem formulados. A percentagem de questões pouco discriminativas, ou seja, com parâmetro de discriminação inferior a 0,3, foi bastante baixa variando entre 3% e 10% nos anos de 2009 e 2010, em que o teste diagnóstico foi realizado em papel, e foi de 0% nos anos seguintes em que o teste foi realizado no computador. Em média a discriminação variou entre 0,887 e 1,205. Concluiu-se, deste modo, que a esmagadora maioria dos itens presentes nos testes diagnósticos eram discriminativos. Os conteúdos matemáticos que revelaram maior capacidade de discriminação dizem respeito a: simplificação de expressões irracionais; gráfico da derivada de uma função e cálculos com logaritmos. Os testes diagnósticos revelaram, portanto, boa qualidade e apresentaram um nível de fiabilidade bastante satisfatório.

Dos conteúdos matemáticos em que os estudantes revelaram maior dificuldade destacaram-se: as inequações envolvendo funções irracionais, logarítmicas e exponenciais; o cálculo de derivadas de funções irracionais e exponenciais e o cálculo de limites com logaritmos. Tal não é surpreendente, uma vez que estes conteúdos são abordados, na sua esmagadora maioria, no 12.^o ano de escolaridade envolvendo, portanto, maior complexidade. Dos conteúdos mais acessíveis para os discentes sobressaíram: as operações com números racionais e as equações do primeiro grau. Estes conteúdos são abordados no 3.^o ciclo do ensino básico pelo que era expectável que os estudantes não sentissem grande dificuldade na sua resolução.

As duas teorias aplicadas, Teoria Clássica de Testes e Teoria de Resposta ao Item, permitiram chegar a resultados e conclusões muito semelhantes.

Relativamente ao banco de questões criado em 2011, quando os testes passaram a ser realizados no computador, verificou-se existir grande heterogeneidade nos conhecimentos inerentes à resolução de alguns dos itens presentes no mesmo grupo de itens. De modo a

que esta situação seja retificada é conveniente que algumas das questões sejam reformuladas para que exista maior homogeneidade nos itens presentes em cada um dos grupos. Pretende-se, desta forma, que todos os estudantes sejam submetidos a questões que têm subjacente à sua resolução o mesmo nível de conhecimentos. Seria também conveniente que existissem mais itens âncora, de modo a permitir a aplicação de um maior número de processos de calibração dos resultados obtidos nos vários anos.

Através de técnicas de calibração, usando Teoria Clássica de Testes e usando Teoria de Resposta ao Item, verificou-se existirem diferenças estatisticamente significativas entre as habilidades médias das diferentes coortes de entrada. Em média o grupo de estudantes que revelou maior habilidade Matemática foi o que realizou o teste diagnóstico aplicado no ano 2009. O grupo de estudantes com menor habilidade média foi o que realizou o teste diagnóstico no ano 2011. Este facto pode dever-se a este ter sido o primeiro ano em que o teste diagnóstico foi realizado no computador e o processo, eventualmente, não ter decorrido da forma mais desejável, a uma alteração da distribuição dos cursos frequentados pelos alunos que realizaram o teste ou a uma efetiva pior preparação a Matemática dos alunos ao nível do ensino básico e secundário. Usando Teoria Clássica de Testes não se verificou a existência de uma tendência linear dos resultados dos estudantes, usando Teoria de Resposta ao Item verificou-se a existência de uma tendência linear negativa dos mesmos. O desvio padrão manteve-se praticamente constante ao longo dos diferentes anos em análise.

Quando comparados diretamente os resultados obtidos nos testes diagnósticos, nos diferentes anos, sem efetuar calibração, verificou-se existir uma tendência linear positiva dos mesmos. A comparação direta dos resultados obtidos nos diferentes anos sugeriu, portanto, que estes estavam a melhorar de ano para ano. A calibração dos resultados é fundamental para que se possa avaliar tendências ao longo do tempo utilizando testes diferentes, pois caso tal não seja feito estamos sujeitos a tirar conclusões erradas.

No que concerne à plataforma online de treino, disponibilizada aos alunos para que pudessem colmatar as dificuldades que evidenciaram, quando realizaram o teste diagnóstico, esta revelou servir os propósitos para que foi criada. Verificou-se existirem evidências, estatisticamente significativas, de que os alunos que treinaram pelo menos duas horas na plataforma obtiveram melhores resultados no teste realizado no âmbito da avaliação da unidade curricular de Cálculo. O treino realizado na plataforma teve, portanto, impacto positivo, estatisticamente significativo, nos resultados obtidos pelos discentes. É importante referir que a não existência de um grupo de estudantes que tenha sido privado de aceder à plataforma constituiu uma limitação para o estudo do impacto do treino na mesma. Indiretamente pode estar a ser medida a motivação dos discentes para melhorar os seus conhecimentos. Para obviar esta questão, seria necessário efetuar um estudo experimental no qual seriam selecionados aleatoriamente os alunos que poderiam aceder à plataforma.

Para que próximas análises relacionadas com os testes diagnósticos sejam facilitadas deve ser melhorada a forma como a informação dos testes diagnósticos e da plataforma de treino é recolhida, uma vez que neste momento esta informação é disponibilizada de forma codificada o que torna bastante morosa a sua análise.

Futuramente poderia ser implementado um banco de itens calibrados sendo os estudantes submetidos a testes diagnósticos de Matemática adaptativos. Neste tipo de testes os alunos vão respondendo a questões mediante a habilidade que vão demonstrando nas suas respostas. Esta opção permitiria diminuir o número de questões a que os alunos têm que responder sem que a fiabilidade dos testes diminuísse.

Bibliografia

- [1] Albano A.D., *equate: An R Package for Observed-Score Linking and Equating*, Oxford, Oxford University Press, 2014.
- [2] Armstrong B.K., White E., Saracci R., *Principles of exposure measurement in epidemiology*, Oxford, Oxford University Press, 1994.
- [3] Baker F., *The Basics of Item Response Theory*, ERIC Clearinghouse on Assessment and Evaluation, University of Maryland, College Park, MD., 2001.
- [4] Bartholomew D.J., *Kendall's Library of Statistics 7: Latent Variable Models and Factor Analysis* (2.nd edition), 1999.
- [5] Bartholomew D.J., Tzamourani P., "The goodness-of-fit of latent trait models in attitude measurement.", *Sociological Methods and Research*, 27, 525-546, 1999.
- [6] Braun H.I., Holland P.W., "Observed-score test equating: A mathematical analysis of some ETS equating procedures.", *Test equating*, New York, 9-49, 1982.
- [7] Brown T.A., *Confirmatory Factor Analysis for Applied Research* (2.nd edition), The Guilford Press, New York, 364-365, 2015.
- [8] Carlson J., Davier M., *Item Response Theory*, ETS R&D Scientific and Policy Contributions Series, 2013.
- [9] Cook L.L., Eignor D.R., Taft H. L., "A comparative study of the effects of recency of instruction on the stability of IRT and conventional item parameter estimates", *Journal of Educational Measurement*, 25, 31-45, 1988.
- [10] Cortina J.M., "What is coefficient alpha? An examination of Theory and Applications", *Journal of Applied Psychology*, vol. 78, 98-104, 1993.
- [11] Cronbach L.J., "Coefficient alpha and the internal structure of tests", *Psychometrika*, vol. 16, №3, 1951.
- [12] De Champlain A.F., "A primer on classical test theory and item response theory for assessments in medical education", *Medical Education*, vol. 44, 109-117, 2010.
- [13] DeVellis R.F., "Classical test theory", *Medical Care* vol. 44, 50-59, 2006.

- [14] Dowling D., Nolan B., *Measuring the effectiveness of a maths learning centre: The Dublin City University experience*. Paper presented at the CETL MSOR Conference, 2006.
- [15] Downing S.M., "Item response theory: applications of modern test theory in medical education", *Medical Education*. vol. 37, 739-745, 2003.
- [16] Ekstrom J., *The Phi-Coefficient, The Tetrachoric Correlation Coefficient, and The Pearson-Yule Debate*. Los Angeles, 2008.
- [17] Embretson S.E., Herschberger S.L. *The new rules of measurement: What every psychologist and educator should know..* Lawrence Erlbaum Associates, Mahwah, New Jersey, 1999.
- [18] Embretson S.E., Reise S.P. *Item response theory for psychologists*. Lawrence Erlbaum Associates, Mahwah, New Jersey, 2000.
- [19] Gill O., Mac An Bhaird C., Ni Fhloinn E., "The Origins, Development & Evaluation of Mathematics Support Services", *Irish Mathematical Society Bulletin*. 66, 51-63, 2010.
- [20] Gill O., O'Donoghue J., "Justifying the Existence of Mathematics Learning Support: Measuring the Effectiveness of a Mathematics Learning Centre", *Proceedings of the ALM*. 14, 2007.
- [21] Gulliksen H., *Theory of mental tests*. New York, 1950.
- [22] Haebara T., "Equating Logistic Ability Scales by a Weighted Least Squares Method", *Japanese Psychological Research*, 22(3), 144-149, 1980.
- [23] Hagenaars J., "Latent Structure Models with Direct Effects Between Indicators" *Sociological Methods & Research*. Sage Publications, Inc., Vol 16, №3, 379-405, 1988.
- [24] Hambleton R.K., Swaminathan H., *Item response theory. Principles and applications*. Boston: Kluwer, 1985.
- [25] Hambleton R.K., Swaminathan H., Rogers J., *Fundamentals of Item Response Theory*. Newbury Park, CA Sage Publications, 1991.
- [26] Hawkes T., Savage M., *Measuring the Mathematics Problem*. London: Engineering Council, 2000.
- [27] Kline T., *Psychological Testing: A Practical Approach to Design and Evaluation*. Sage Publications, London, 2005.
- [28] Kolen M.J., Brennan R.L., *Test equating, scaling and linking: methods and practices*. Springer, New York, 2004.
- [29] Lee S., Harrison M., Pell P., Robinson C. "Predicting performance of first year engineering students and the importance of assessment tools therein" *Engineering Education*. 3, 44-51, 2008.

- [30] Lehman E.L., Casella G., *Theory of Point Estimation*. (2nd edition), 1998.
- [31] Little T.D., *The Oxford Handbook of Quantitative Methods, Volume 1: Foundations*. Oxford University Press, New Jersey, 2013.
- [32] Livingston S.A., *Equating test scores (without IRT)*. Educational Testing Service, New Jersey, 2004.
- [33] Lord F.M., Novick M.R. *Statistical theories of mental test scores*. Reading MA: Addison Welsley Publishing Company, 1968.
- [34] Loyd B.H., Hoover H.D., "Vertical Equating Using the Rasch Model." *Journal of Educational Measurement*, 17(3), 179-193, 1980.
- [35] LTSN., *Diagnostic Testing for Mathematics*. UK: Learning and Teaching Support Network Maths Team Project., 2002.
- [36] Mac An Bhaird C., O'Shea A. "Is Mathematics Support worthwhile? An overview of the 3rd Irish Workshop on Mathematics Learning and Support Centres" *MSOR Connections*. 9, 52-59, 2009.
- [37] Marco G.L., "Item Characteristic Curve Solutions to Three Intractable Testing Problems.", *Journal of Educational Measurement*, 14(2), 139-160, 1977.
- [38] McDonald M.E., *Systematic Assessment of Learning Outcomes: Developing Multiple-choice Exams*. Sudbury, 2002.
- [39] McGillivray H., *Learning Support in Mathematics and Statistics in Australian Universities: A guide for the university sector*. Sidney: Australian Teaching and Learning Council, 2008.
- [40] Meijer R.R., Sijtsma K., *Methodology review: evaluating person fit..* Applied Psychological Measurement 25(2), 107-135, 2001.
- [41] Miller M.D., Linn R.L., "Invariance of item characteristic functions with variations in instructional coverage" *Journal of Educational Measurement*. 25, 205-219, 1988.
- [42] Pasquali L., *Psicometria: Teoria e aplicações*. Brasília, Editora da Universidade de Brasília, 1997.
- [43] Rasch G., *Probabilistic Models for Some Intelligence and Attainment Tests*, Copanha-gen: Denmark's Paedagogiske Institut, 1960.
- [44] Rijmen F., Tuerlinckx F., De Boeck P., Kuppens P., "A Nonlinear Mixed Model Framework for Item Response Theory" *Psychological Methods*, Vol. 8, №2, 185-205, 2003.
- [45] Rocha C., *O ensino da Matemática no IST*, Lisboa, 2002.
- [46] Savage M. D., *Tackling the Maths Problem: Is it More Extensive than We Thought? Are Solutions in Sight?*. Paper presented at the Mathematical Education of Engineers IV, IMA Conference, 2003.

- [47] Sheridan B., "How Much Do Our Incoming First Year Students Know?: Diagnostic Testing in Mathematics at Third Level" *Irish Journal of Academic Practice*. Vol. 2: Iss. 1, Article 3, 2013.
- [48] Sik-Yum L., *Handbook of Latent Variable and Related Models*. (1.st Edition) Elsevier B.V., Oxford, 135-161, 2007.
- [49] Smith A., *Making Mathematics Count: The report of Professor Adrian Smiths Inquiry into Post-14 Mathematics Education*. London, 2004.
- [50] Spearman C., *The proof and measurement of association between two things*. Am J Psychol, 3:1-18, 1904.
- [51] Steven P.R., "A Comparison of Item- and Person-Fit Methods of Assessing Model-Data Fit in IRT" *Applied Psychological Measurement*. Vol. 14, №2, 127-137, 1990.
- [52] Stocking M.L., Lord F.M., "Developing a Common Metric in Item Response Theory.", *Applied Psychological Measurement*, 7(2), 201-210, 1983.
- [53] Strauss M.E., Smith G.T., "Construct validity: Advances in theory and methodology" *Annual Review of Clinical Psychology*. 5, 1-25, 2009.
- [54] Streiner D.L, Norman G.R. *Health measurement scales*. Oxford University Press, New York, 213-226, 2003.
- [55] Tarkkonen L., Vehkalahti K., "Measurement errors in multivariate measurement scales" *Journal of Multivariate Analysis*. 96, 172-189, 2004.
- [56] Terlow C., de Goede R., Keinhuis M., *The effectiveness of a special mathematics course for improving the transition from secondary education to higher professional education*. Paper presented at the 7th International conference of The London Scholarship of Teaching and Learning, 2008.
- [57] Thorpe G., Favia A., "Data Analysis Using Item Response Theory Methodology: An Introduction to Selected Programs and Applications." *Psychology Faculty Scholarship*. Paper 20, 2012.
- [58] Traub R.E., "Classical Test Theory in Historical Perspective" *Educational Measurement: Issues and Practice*. Ontario, 1997.
- [59] Varma, S., *Preliminary Item Statistics using Point-Biserial Correlation and P-Values*. Educational Data Systems, Inc., Morgan Hill, 2006.
- [60] Weeks J.P., *plink: An R Package for Linking Mixed-Format Tests Using IRT - Based Methods*. Journal of Statistical Software, vol. 35, 2010.
- [61] Wilks S.S., "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses" *The Annals of Mathematical Statistics*. 9, 60-62, 1938.

Apêndice A

Teste Diagnóstico 2010



FC FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO
Departamento Matemática



Teste diagnóstico 2010

Aferição de conhecimentos em Matemática básica

O principal objectivo do teste é aferir carências de carácter elementar na formação em Matemática e não se destina a fazer qualquer tipo de avaliação individual.

Data: 22 de Setembro de 2010 (quarta-feira)

Hora de início: 14h

Locais: salas a indicar

Duração: 1h30m (sem tolerância)

Indicações:

- O teste é constituído por **7 (sete) folhas** e tem **30 (trinta) questões** de escolha múltipla, com 4 opções cada, das quais uma e apenas uma está correcta.
- A opção escolhida para cada questão deve ser indicada no formulário distribuído à parte. Neste formulário tem que indicar o seu número mecanográfico, de acordo com as instruções que receberá no início do teste.
- Use o verso das folhas para rascunho.
- Não é permitido o uso de máquina de calcular.

Não esqueça: as respostas são indicadas no formulário distribuído à parte.

► **Problema 1** Se x é um número real diferente de 0, então a expressão

$$\left(\frac{2}{x} + \frac{x}{2}\right)^2 - \left(\frac{2}{x} - \frac{x}{2}\right)^2$$

é igual a:

- A. $\frac{2}{x}$ B. x C. 4 D. 0
-

► **Problema 2** A figura representa um hexágono regular inscrito num círculo de raio igual a 1

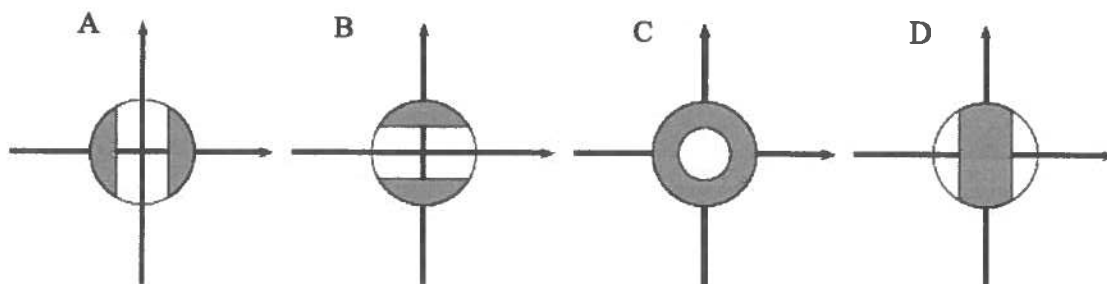


A área do hexágono é igual a:

- A. $\frac{3}{2}$ B. $3\sqrt{2}$ C. $\frac{3\sqrt{3}}{4}$ D. $\frac{3\sqrt{3}}{2}$
-

► **Problema 3** Qual das zonas sombreadas poderá representar a região do plano definida por

$$\{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 4 \wedge x^2 \geq 1\} ?$$



- **Problema 4** Se $x > 10$, então

$$|3 - |5 - x||$$

é igual a:

- A. $x - 8$ B. $8 - x$ C. $-2 + x$ D. $x + 2$
-

- **Problema 5** Se α é um ângulo do 4º quadrante e $\cos \alpha = \frac{1}{3}$, então $\sin \alpha$ é igual a:

- A. $\frac{\sqrt{2}}{3}$ B. $\frac{2\sqrt{2}}{3}$ C. $-\frac{2\sqrt{2}}{3}$ D. $-\frac{\sqrt{2}}{3}$
-

- **Problema 6** A equação da recta que passa no ponto $(1, 1)$ e é perpendicular à recta de equação $y = mx + k$, com $m \neq 0$, é:

- A. $y = -\frac{x}{m} - 1 + \frac{1}{m}$ B. $y = -\frac{x}{m} + 1 + \frac{1}{m}$
 C. $y = \frac{x}{m} - 1 - \frac{1}{m}$ D. $y = \frac{x}{m} + 1 - \frac{1}{m}$
-

- **Problema 7** O conjunto solução da inequação

$$|3 + 2x| \geq 1$$

é:

- A. $] -\infty, -2] \cup [-1, +\infty[$ B. $] -\infty, 1] \cup [2, +\infty[$
 C. $[-2, -1]$ D. $] -\infty, -2] \cup [-1, 2]$
-

- **Problema 8** Se $x > 0$, a expressão

$$\sqrt{x} + \frac{1}{2} \frac{x}{\sqrt{x}}$$

é igual a:

- A. $\frac{x}{\sqrt{x}}$ B. $\frac{x+1}{2\sqrt{x}}$ C. $\frac{3}{2}\sqrt{x}$ D. $\frac{3}{2}x$

- **Problema 9** Se $x \geq 3$, a expressão

$$\sqrt{x^2 - 9}$$

é igual a:

- A. $x - 3$ B. $|x| - 3$
C. $\sqrt{x - 3}\sqrt{x + 3}$ D. $|x - 3|$
-

- **Problema 10** No seu domínio, a expressão

$$e^{5 \ln(xy)}$$

é igual a:

- A. $e^5 xy$ B. $x^5 y^5$ C. $e^5(x + y)$ D. $5xy$
-

- **Problema 11** A inequação $\frac{x}{x+1} > 1$ é

- A. impossível B. equivalente a $x > x + 1$
C. equivalente a $x < x + 1$ D. equivalente a $x < -1$
-

- **Problema 12** Qual o valor de n se

$$\frac{2^{500}}{2^{2n-1}} = \frac{1}{32} \quad ?$$

- A. $n = 253$ B. $n = 258$ C. $n = 523$ D. $n = 250$
-

- **Problema 13** Considere as igualdades seguintes

I. $\log(\sqrt{a}) = \sqrt{\log a}$ II. $\log\left(\frac{a}{b}\right) = \frac{\log a}{\log b}$

III. $\log(a + b) = \log a + \log b$ IV. $\log a + \log b = \log(ab)$

Quanto à sua validade $\forall a, b > 0$, podemos afirmar que:

- A. I é verdadeira B. II é verdadeira
C. III é verdadeira D. IV é verdadeira

► **Problema 14** Se $f(x) = \frac{1}{x}$ e $g(x) = e^x + 1$, então $(f \circ g)(x)$ é igual a:

- A. $e^{1/x} + 1$ B. $\frac{1}{e^x} + 1$ C. $\frac{1}{e^x + 1}$ D. $\frac{1}{e^{-x} + 1}$
-

► **Problema 15** Se $f(x) = 1 + \sqrt{2 + 2x}$ então, $\forall x > -1$, $f'(x)$ é igual a:

- A. $(2 + 2x)^{-1/2}$ B. $1 + (2 + 2x)^{-1/2}$
 C. $\frac{1}{2}(2 + 2x)^{-1/2}$ D. $1 + \frac{1}{2}(2 + 2x)^{-1/2}$
-

► **Problema 16** Se $f(x) = x e^{x^2}$, então $f'(x)$ é igual a:

- A. e^{x^2} B. $e^{x^2} + x e^{x^2}$
 C. $e^{x^2} + 2x^2 e^{x^2}$ D. $e^{x^2} + 2x e^{x^2}$
-

► **Problema 17** O limite

$$\lim_{x \rightarrow \infty} x \ln(x^2 + 1)$$

é igual a:

- A. $+\infty$ B. $-\infty$ C. 0 D. indeterminação
-

► **Problema 18** O limite

$$\lim_{x \rightarrow +\infty} \frac{\sqrt{x}}{x}$$

é igual a:

- A. 1 B. $+\infty$ C. 0 D. $\frac{1}{2}$

► **Problema 19** Se $(-1+x)\left(2-\frac{1}{x}\right) = 1$, então

A. $x = 2 \vee x = 1$

B. $x = \frac{1+\sqrt{2}}{2} \vee x = \frac{1-\sqrt{2}}{2}$

C. $x = \frac{2+\sqrt{2}}{2} \vee x = \frac{2-\sqrt{2}}{2}$

D. $x = 2 \vee x = \frac{1}{2}$

► **Problema 20** A função $f: \mathbb{R} \rightarrow \mathbb{R}$ definida por $f(x) = x^2 + bx + c$, com $b, c \in \mathbb{R}$

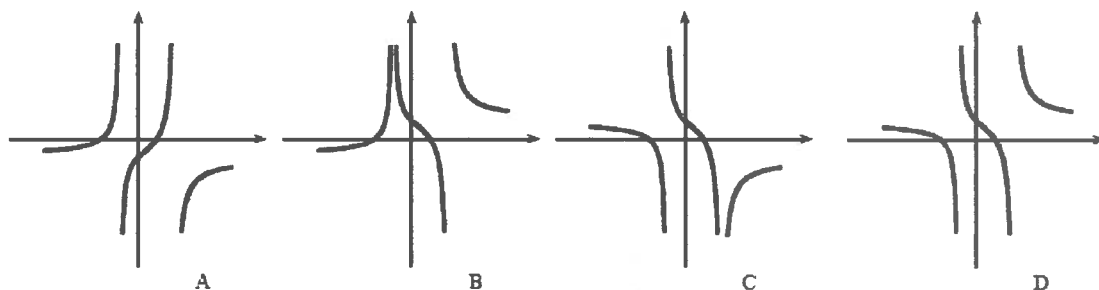
A. é decrescente.

B. tem inversa.

C. não é injectiva.

D. tem o contradomínio contido em $] -\infty, m]$ para algum $m \in \mathbb{R}$.

► **Problema 21** O gráfico da função $f(x) = \frac{(x-1)(x+2)}{(x+1)(x-2)}$ tem o aspecto seguinte:



► **Problema 22** Considere a função $f(x) = \frac{1}{x}$, $x \neq 0$. Das seguintes afirmações

1. $f(ab) = f(a)f(b)$

2. $f(a^2) = (f(a))^2$

3. $f(a+b) = f(a) + f(b)$

4. $f\left(\frac{1}{a}\right) = \frac{1}{f(a)}$

quantas são verdadeiras, $\forall a, b \in \mathbb{R} - \{0\}$?

A. exactamente uma

B. exactamente duas

C. exactamente três

D. todas

- **Problema 23** Se $f(x) = x - 2$ e $g(x) = 2x + 1$ então

$$(f \circ g)(2) - (g \circ f)(-1)$$

é igual a:

- A. 4 B. 8 C. -4 D. -8

- **Problema 24** O domínio da função

$$f(x) = \frac{\sqrt{x^2 - 1}}{|x| - 1}$$

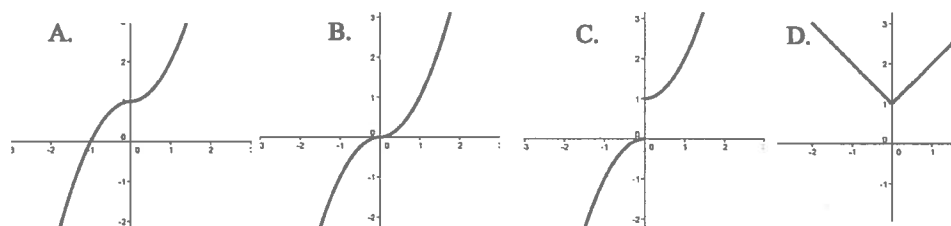
é:

- A. $[-1, 1]$ B. $] -\infty, -1] \cup [1, +\infty[$
 C. $] -\infty, -1[\cup]1, +\infty[$ D. $\mathbb{R} \setminus \{\pm 1\}$

- **Problema 25** Se $f : \mathbb{R} \rightarrow \mathbb{R}$ é uma função tal que $f(0) = 1$ e, $\forall x \in \mathbb{R}$,

$$f'(x) = 2|x|,$$

então o gráfico de f poderá ser:



- **Problema 26** Os gráficos das funções $f(x) = \sin x$ e $g(x) = \cos x$:

- A. intersectam-se num único ponto.
 B. não se intersectam.
 C. intersectam-se numa infinidade de pontos.
 D. intersectam-se exactamente em dois pontos.

- **Problema 27** Se $x \in \mathbb{R}$ então a expressão

$$\sin\left(\frac{\pi}{2} + x\right) - \cos(x + \pi)$$

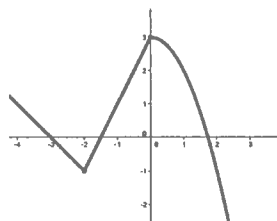
é igual a:

- A. $\cos x - \sin x$ B. $2 \cos x$
C. 0 D. $\cos x + \sin x$

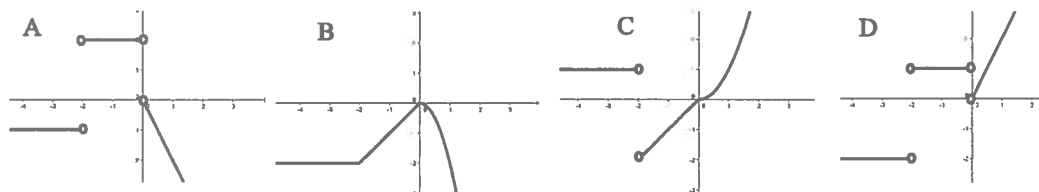
- **Problema 28** A função $h(x) = 1 + 2 \sin x$ tem, no intervalo $]0, 2\pi[$,

- A. nenhum extremo B. exactamente 1 extremo
C. exactamente 2 extremos D. 3 extremos

- **Problema 29** A função cujo gráfico tem o aspecto seguinte



poderá ter como função derivada:



- **Problema 30** Se $f: \mathbb{R} \rightarrow \mathbb{R}$ é a função definida por $f(x) = 2x + 3$, então:

- A. f não tem inversa
B. f tem inversa e $f^{-1}(x) = \frac{1}{2x+3}$
C. f tem inversa e $f^{-1}(x) = \frac{x-3}{2}$
D. f tem inversa e $f^{-1}(x) = \frac{2}{x} - 3$

***** FIM *****