Resultados do **exame-piloto**

Uma análise integral sobre as novas questões para as provas das certificações de distribuição da ANBIMA

CERTIFICAÇÕES ANBIMA

Seu passaporte para o mercado financeiro

ANBIMA

Sumário

Por que testamos nossos modelos de questões 2

Planejando o exame-piloto 4

Plano de análise estatística 6

Principais resultados 8

Estrutura das novas provas ANBIMA 13

Considerações finais 15

Referências 16

Por que testamos nossos modelos de questões

Em abril de 2024, anunciamos uma mudança importante: as certificações de distribuição **CPA-10, CPA-20 e CEA** seriam descontinuadas em 2026, dando lugar a três novas certificações de distribuição da ANBIMA — **CPA, C-Pro R e C-Pro I**. Essa transformação foi fruto de intensas mudanças no mercado financeiro e de capitais brasileiro, que exigiram uma atualização profunda no nosso modelo de avaliação.

Por isso, para criar certificações que de fato refletissem as necessidades do mercado, conduzimos em parceria com a Deloitte um amplo estudo para analisar as melhores práticas em certificações e entender as necessidades reais do nosso mercado (estudo completo: **"Um olhar sobre as certificações"**). O objetivo era claro: criar certificações alinhadas aos perfis de profissionais que atuam no setor e às expectativas das instituições.

Um dos principais aprendizados foi que, além das **habilidades técnicas**, os profissionais precisam demonstrar **habilidades comportamentais**. Ou seja, percebemos que era necessário avaliar a capacidade de aplicar conhecimentos técnicos de forma ética, colaborativa e com foco em resultado no dia a dia de trabalho.

Para isso, desenvolvemos um novo modelo avaliativo. **Afinal, um novo jeito de avaliar exige um novo jeito de perguntar.** Reformulamos a matriz de avaliação, criamos modelos de questões e construímos um banco inédito de perguntas, que começou a ser testado em exame-piloto com participação voluntária de profissionais do mercado.

Esses testes foram fundamentais para garantir a validade e confiabilidade das provas, além de assegurar que manteremos os padrões de equidade e transparência que direcionam todas as certificações ANBIMA. Também aproveitamos para calibrar elementos como **tempo de prova**, **número de questões** e **grau de dificuldade**, visando um equilíbrio que respeite o perfil de cada certificação.

Além da parte técnica, o exame-piloto teve um papel fundamental na aproximação com o mercado, pois permitiu que profissionais, escolas preparatórias e RHs de instituições associadas conhecessem de perto o novo modelo. Essa transparência faz parte da nossa premissa desde o início do projeto: garantir que todos os públicos tenham acesso às etapas que conduzirão às novas provas em fevereiro de 2026.

Com esse exame-piloto, buscamos responder a perguntas essenciais:

- Os modelos de questões de fato avaliam o que se pretende, com base nos critérios da matriz de avaliação ANBIMA?
- É viável levar modelos mais inovadores de questões para certificações desse porte?
- O grau de dificuldade das questões está adequado a cada perfil de profissional?
- Quantas questões deve ter cada uma das provas?
- Qual o tempo de prova adequado para cada uma das certificações?

Este documento resume as respostas que obtivemos a essas dúvidas.

Planejando o exame-piloto

A aplicação do exame-piloto aconteceu entre os dias 26 e 31 de março de 2024 e contou com 185 participantes, distribuídos de forma presencial e remota.

Foram 53 participantes no formato remoto, avaliando não só o exame, mas também a experiência de realizar uma prova fora dos centros de testes; e 132 participantes nos centros de testes de São Paulo e Santo André.

Os critérios definidos para aplicação do exame-piloto e avaliação dos resultados para melhor definição e encaminhamento das novas provas foram:

- Uma prova por certificação;
- 60 questões por prova;
- Questões distribuídas em graus de dificuldade;
- Questões distribuídas por níveis cognitivos (compreensão, aplicação e análise);
- Duração máxima de 180 minutos para a realização da prova.

Em relação ao banco de questões, a amostra para o exame-piloto se baseou nos tipos apresentados anteriormente no relatório **Nosso jeito de avaliar:**

- **Questões de árvore de decisão:** consistem na simulação de uma conversa entre profissional e cliente, na qual o candidato precisa demonstrar ter conhecimento técnico sobre o tema da questão e também **habilidades comportamentais** necessárias para exercer suas atividades profissionais. Esse formato esteve presente nas certificações CPA e C-Pro R.

escolha, e estão sempre associadas a um case ou minicase, que se relaciona com a prática profissional. A nomenclatura "dissertativa de resposta curta" foi escolhida porque esse modelo pressupõe respostas objetivas, que podem ser dadas por meio de uma frase simples, um número ou mesmo uma palavra.

- **Questões de múltipla escolha contextualizada:** predominantes em todas as certificações, as questões de múltipla escolha avaliam habilidades de caráter mais técnico, sempre associadas a um contexto, que se relaciona com a prática profissional;

- **Cases:** trazem como raiz uma situação específica, disparando um desafio a ser resolvido pelo profissional. As situações estão acompanhadas de até quatro questões de múltipla escolha ou dissertativa de resposta curta. Os cases estiveram nos exames para C-Pro R e C-Pro I;

- **Minicases:** dinâmica similar aos cases, porém com situações mais curtas e menos complexas e limite de até duas questões. Os minicases estiveram presentes nos exames para C-Pro R e C-Pro I;

- **Dissertativas de resposta curta:** avaliam habilidades de caráter mais técnico, assim como as de múltipla

Plano de análise estatística

Os resultados foram verificados com base na **metodologia de análise quantitativa/estatística**, de modo a produzir resultados do ponto de vista psicométrico.

Uma vez realizadas as provas, a avaliação da qualidade dos testes e das questões foi feita a partir da Teoria de Resposta ao Item (TRI) e não pela Teoria Clássica de Testes (TCT). A preferência pela TRI aconteceu em especial pela necessidade de **parâmetros de comparabilidade** entre os testes. A partir da TCT não poderíamos garantir que dois exames tivessem o mesmo grau de dificuldade e, portanto, que o critério de atribuição de certificado fosse isométrico.

Com a TRI, essa unicidade do critério de aprovação passa a ser possível. A opção metodológica entre as diversas abordagens contempladas pela TRI foi pelo **modelo de três parâmetros**, descrito em Birnbaum (1968), a partir do qual calculamos:

- capacidade de discriminação, utilizado como indicador de qualidade;
- dificuldade;
- probabilidade de acerto casual (chute).

O **parâmetro de discriminação** foi utilizado como critério de qualidade das questões, a partir do qual foram definidas as questões que devem ser submetidas à revisão ou descarte e as que devem ser preferidas na construção de provas.

A referência para definição do nível de qualidade das questões segue o padrão estabelecido por Baker (2001), descrito a seguir:

Intervalo	Conceito
a = 0	Nenhuma
0 < a < 0,35	Muito baixa
0,35 ≤ a < 0,65	Baixa
0,65 ≤ a < 1,35	Moderada
1,35 ≤ a < 1,70	Alta
1,70 ≤ a	Muito alta
a > ∞	Perfeita

Tabela 1: níveis de qualidade a partir do parâmetro de discriminação

Outros indicadores, incluindo a **correlação bisserial**, comum à TCT, foram utilizados para análise da qualidade das questões, mas os parâmetros calculados pela TRI apresentaram maior precisão (Orlando, 2000) além de permitirem **comparabilidade em contextos distintos de aplicação.**

Considerada a necessidade de ajustar a dificuldade das questões ao critério de aprovação, foram calculadas também as Curvas de Informação dos Itens (CII). A CII ilustra a porção do domínio de desempenho em que a questão contribui com informações adicionais. Por exemplo: uma questão fácil não serve para discriminar candidatos de alto desempenho. É a partir das CIIs das questões que se define a Curva de Informação do Teste (CIT) e o domínio de desempenho da prova como um todo. Como indicado em Baker (2001, p.119), o ideal é que a CIT tenha uma distribuição normal com pico coincidente com o critério de aprovação no exame de certificação. A adequação da CIT do exame à referência foi aferida no exame-piloto e poderá ser aprimorada com a seleção de questões de parâmetros conhecidos.

Outros indicadores como o coeficiente alfa ("de Cronbach") e o índice de separação foram calculados para análise da qualidade dos testes em um indicador único. As metodologias para cálculo dos dois indicadores são mais complexas e, portanto, não são reproduzidas aqui, mas podem ser compreendidas a partir de Cronbach (1951), Novick (1967), Mallison (2004) e Shojima (2002).

Principais resultados encontrados

A partir dos resultados obtidos no exame-piloto, é possível afirmar que estamos na direção certa com nossos modelos de questões e que, com pequenos ajustes nos parâmetros de construção do nosso banco, teremos em 2026 provas de certificação válidas e confiáveis do ponto de vista avaliativo.

4.1 Distribuição da qualidade e nível de dificuldade das provas

De modo geral, as questões utilizadas no exame-piloto foram avaliadas em termos de qualidade, validade avaliativa e grau de dificuldade, conforme gráfico a seguir:

Qualidade dos itens | Formato

Distribuição da dificuldade empírica, qualidade, validade e quantidade de itens por formato de contexto e de resposta

Formato de contexto	Formato de resposta	Dificuldade do item (%) (Fácil / Médio / Difícil)	Qualidade do item (%) (Excelente / Bom / Adequado / Inadequado / Contraproducente)	Item válido (%) (Verdadeiro / Falso)
Simples	Múltipla escolha	17 / 40 / 43	2 / 2 / 20 / 23 / 53	4 / 96
Árvore de diálogo	Árvore de diálogo	100	25 / 50 / 25	25 / 75
	Múltipla escolha	38 / 63	13 / 38 / 50	100
Case	Dissertativa de resposta curta	83 / 17	33 / 67	100
	Múltipla escolha	20 / 60 / 20	7 / 7 / 7 / 27 / 53	13 / 87
Minicase	Dissertativa de resposta curta	50 / 38 / 13	38 / 63	100
	Múltipla escolha	100	100	100

Ou seja, podemos afirmar sobre a qualidade do banco de questões testado que:

- a maioria das questões avaliadas apresentou **qualidade adequada**, boa ou excelente, sendo que as que foram avaliadas como inadequadas ou contraproducentes foram excluídas do nosso banco-modelo;
- o maior número de questões contraproducentes está nas certificações C-Pro I e CPA, o que aponta para uma revisão de como definimos os critérios para essas certificações;
- em termos de qualidade e validade das questões, a C-Pro R foi considerada a prova mais bem construída.

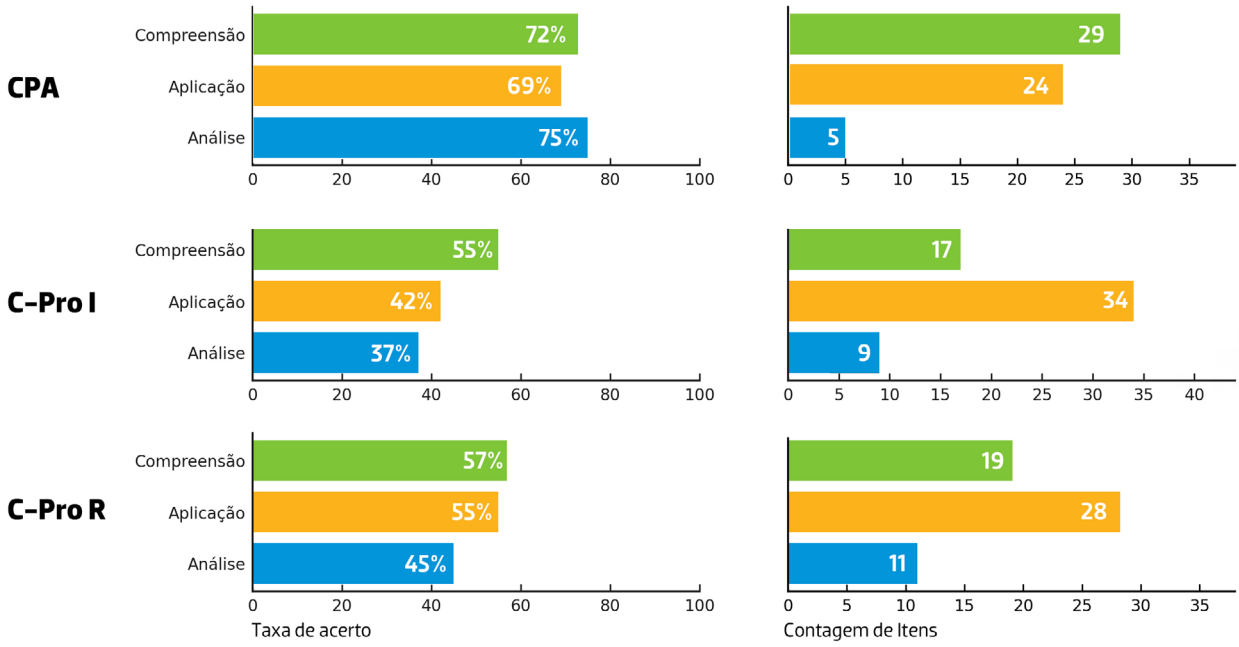
O gráfico nos informa ainda sobre o grau de dificuldade das questões, sendo que:

- a CPA foi a prova que concentrou o maior número de questões fáceis;
- A C-Pro I foi a prova que concentrou o maior número de questões difíceis;
- A C-Pro R foi a prova mais balanceada em termos de nível de dificuldade.

4.2 Análise das questões por atributos da matriz de avaliação ANBIMA

A partir dos atributos definidos na matriz de avaliação das novas certificações, analisamos as questões no exame-piloto a partir do nível cognitivo, dos macrotemas presentes nos Programas Detalhados (PDs) de cada certificação e da habilidade comportamental envolvida no contexto da questão.

Desempenho por nível cognitivo

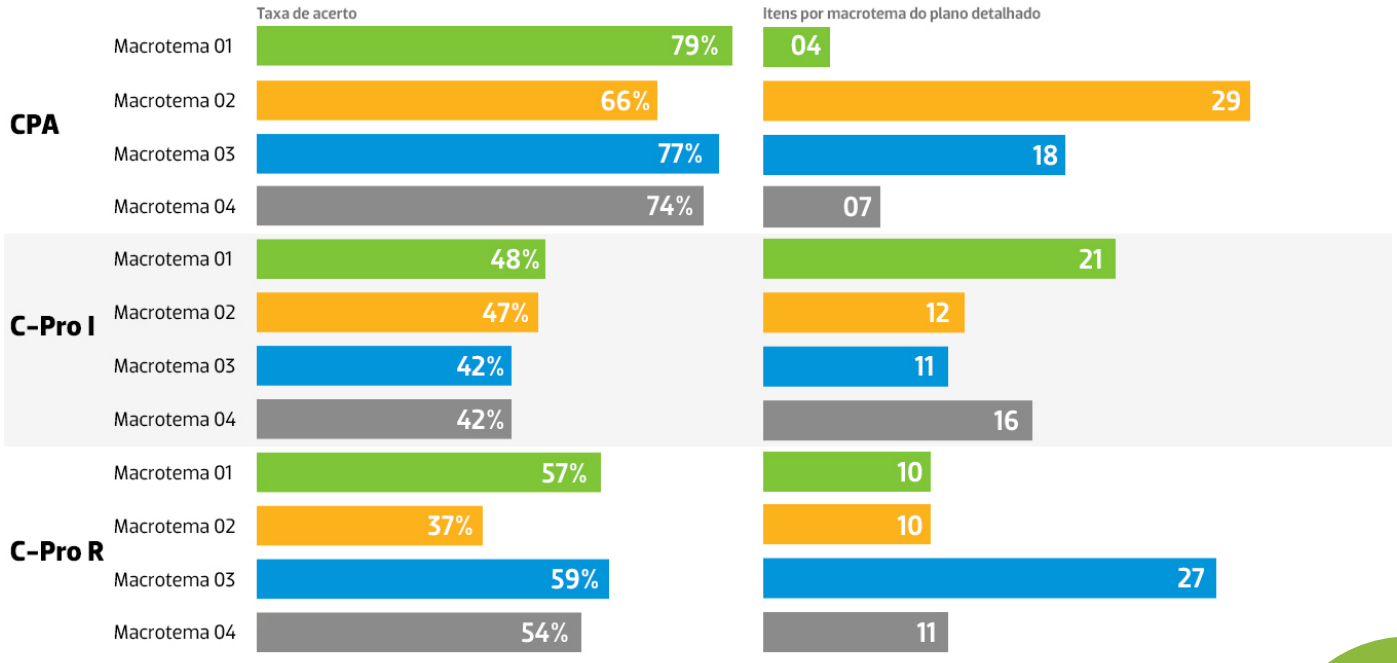


O gráfico anterior nos mostra que:

- a CPA e a C-Pro R tiveram mais uniformidade na taxa de acerto em relação aos três níveis cognitivos, o que pode ser explicado por terem questões mais fáceis (no caso da CPA) e mais balanceamento no grau de dificuldade (no caso da C-Pro R);
- a C-Pro I, por sua vez, apresentou uma curva de acerto dentro do esperado para uma prova de nível difícil, uma vez que essa curva caiu com o aumento do nível cognitivo das questões. Ou seja, nas questões de nível mais complexo, a taxa de acerto foi menor.

Em relação aos macrotemas dos PDs a que as questões se referiam, pudemos observar que, de modo geral, as taxas de acerto tiveram pouca variação nas três provas:

Desempenho por macrotema



É importante destacar, como exceção, o macrotema 2 da C-Pro R, Indicação de investimentos, que apresentou uma taxa de acerto acentuadamente menor do que nos outros três macrotemas do programa.

Aqui, vale pontuar que o ato de indicar investimentos é o centro da atividade do profissional C-Pro R, uma das principais mudanças que inserimos em nosso modelo avaliativo. Isso nos sinaliza que, de fato, os profissionais precisam estar mais bem preparados em relação a questões que avaliam por meio de contextos o seu dia a dia profissional, já que tiveram uma taxa menor de acerto nesse macrotema.

As **habilidades comportamentais** foram avaliadas no contexto das questões, como previsto no documento **Novo Jeito de Avaliar.** Ou seja, para cada questão foi avaliado o grau de inserção de cada habilidade em seu contexto.

Desempenho por habilidades comportamentais

Taxa de acerto por habilidades comportamentais e nível de relacionamento

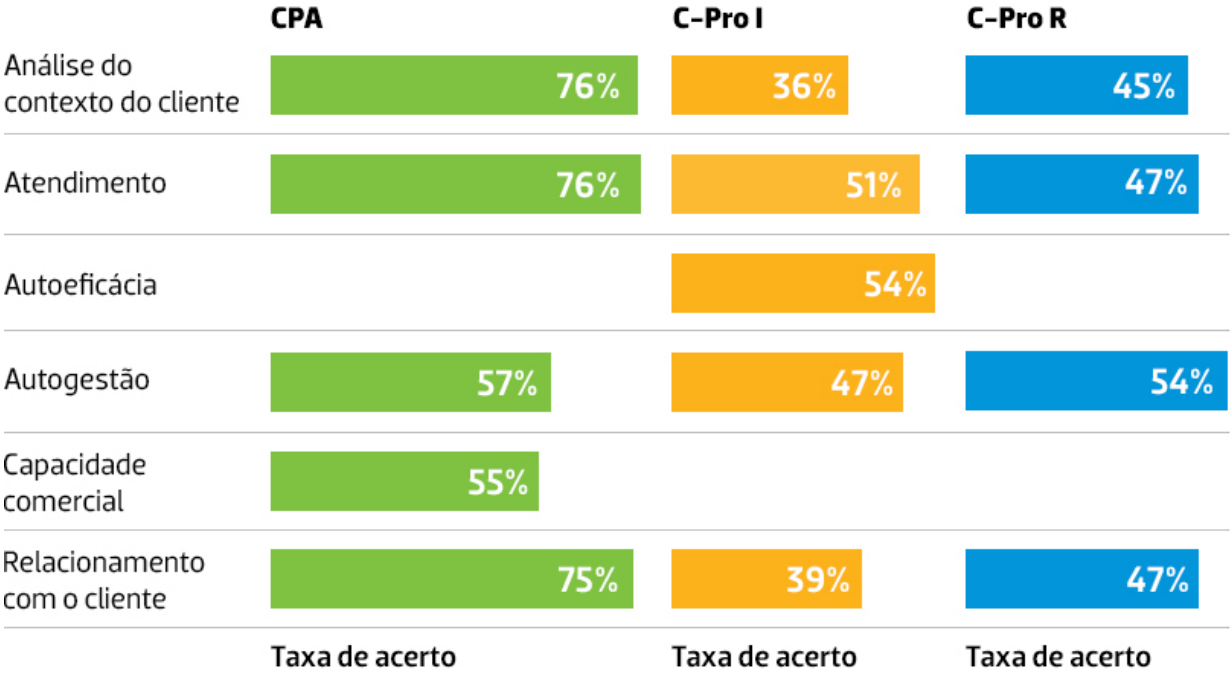
	01 Não está envolvida no contexto	02 É periférica no contexto	03 É importante para o contexto	04 É estruturante para o contexto
Análise do contexto do cliente		52%	58%	47%
Atendimento	53%	47%	56%	65%
Autoeficácia	67%	52%	60%	54%
Autogestão		59%	55%	52%
Capacidade comercial	54%	49%	66%	55%
Comunicação interpessoal	56%	52%	58%	
Persuasão	53%	52%	63%	
Relacionamento com o cliente	56%	50%	58%	54%
Suporte social	52%	54%	61%	

O gráfico anterior mostra que, de modo geral, alcançamos uma distribuição uniforme das habilidades nos contextos, uma vez que **comunicação interpessoal, persuasão e suporte social,** que não aparecem como estruturantes dos contextos, de fato não são centrais para as atividades específicas dos profissionais. Por isso, é compreensível que se mostrem de forma menos centralizada.

Por fim, em termos de **habilidades comportamentais** que estruturam o contexto das questões, tivemos os seguintes resultados:

Desempenho por habilidade comportamental e certificação

Taxa de acerto por habilidade comportamental e nível de relacionamento



O gráfico evidencia alguns pontos interessantes sobre o direcionamento das questões **(habilidades técnicas e comportamentais)** em relação ao perfil de cada certificação:

- A C-Pro I é uma prova que está bem construída nesse aspecto, uma vez que a única habilidade comportamental que não estruturou seus contextos de questões foi capacidade comercial. Isso está muito alinhado ao perfil da certificação, que não prevê um profissional que lida diretamente com o cliente.
- Por outro lado, na CPA, não temos uma concentração de questões estruturadas por autoeficácia, o que também se alinha ao perfil de um profissional que está entrando no mercado.
- Já na C-Pro R, como ponto de atenção, está a falta de questões estruturadas por capacidade comercial, que deve ser sim um atributo do profissional certificado por essa prova.

Estrutura das novas provas ANBIMA

Além do desempenho das questões de prova cruzado com parâmetros apresentados aqui, foram analisados também pontos como:

- desempenho do **candidato x tempo de resolução** da prova;
- desempenho do **candidato x formato da questão;**
- desempenho do **candidato x posição da questão** na prova (início, meio, fim) / fadiga.

Com base em todos esses critérios, concluímos que a estrutura que melhor atende aos requisitos das novas certificações para a estruturação das novas provas é:

Certificação	Total de questões	Modelo de questões	Grau de dificuldade	Tempo de prova
CPA	50	40 questões de múltipla escolha contextualizada 10 questões relacionadas à arvore de decisão	25% 50% 25%	2h30m
C-Pro R	45	30 questões de múltipla escolha contextualizada 15 questões relacionadas a case e/ou árvore de decisão	25% 50% 25%	2h30m
C-Pro I	40	30 questões de múltipla escolha contextualizada 10 questões relacionadas a case	25% 50% 25%	2h30m

Fácil Médio Difícil

É importante ressaltar que:

- o tempo de prova necessário foi definido com base no tempo médio de resposta de cada tipo de questão, sendo que a quantidade e modelo por prova foi ajustado de acordo com o perfil da certificação, respeitando o tempo máximo de 2h30;
- o minicase foi excluído do nosso modelo de avaliação, uma vez ter demonstrado um baixo índice de discriminação nas provas, em relação às questões de múltipla escolha contextualizada e os cases;
- a árvore de decisões se divide em cinco etapas de diálogo e duas questões de múltipla escolha;
- o case se divide em duas questões de múltipla escolha ou uma questão de múltipla escolha seguida de uma questão dissertativa.

Considerações finais

O exame-piloto realizado pela ANBIMA marcou mais do que uma etapa técnica do processo de transição das certificações de distribuição, pois evidenciou um exercício de escuta, transparência e construção coletiva. Em um cenário de mudanças profundas no mercado financeiro, testar previamente os novos formatos de avaliação é essencial para assegurar a idoneidade, a confiabilidade e a validade das futuras provas.

Esses testes não apenas garantem **rigor estatístico** ao processo, como também reafirmam o compromisso da ANBIMA com uma certificação que seja justa e alinhada às competências exigidas pelo mercado atual. Ao convidar voluntariamente pessoas do setor para participarem do exame-piloto, colocamos o próprio público-alvo no centro da análise – afinal, não basta mudar a forma de perguntar: é preciso entender como essa mudança repercute na prática e na experiência real de quem será avaliado.

Esse processo de escuta ativa nos permitiu **calibrar**, **ajustar** e, principalmente, **reforçar** que a nova matriz de avaliação não é um projeto isolado. Pelo contrário, é uma jornada construída a muitas mãos, que envolve educadores, instituições, profissionais do mercado, especialistas em avaliação e equipes técnicas da ANBIMA. Todos contribuíram para desenhar uma certificação mais conectada com os desafios reais da profissão, mais sensível às **habilidades comportamentais** e mais coerente com as diferentes fases da trajetória de um profissional do setor financeiro.

Os principais resultados obtidos com o exame-piloto – como a validação dos modelos de questões, o equilíbrio no grau de dificuldade das provas e a definição de estruturas adequadas para cada certificação – fornecem os alicerces para que as novas provas ganhem contornos mais consistentes, realistas e eficazes. É importante lembrar que estamos diante de um processo em constante evolução. A partir dessas primeiras respostas, novas perguntas surgirão, novas análises serão feitas e novos ajustes serão incorporados. Assim como o mercado se transforma, nosso modo de avaliar também deve acompanhar esse movimento, sempre de forma ética, colaborativa e com foco no desenvolvimento dos profissionais.

A ANBIMA agradece a todos que contribuíram para essa construção.

Referências

Akour, M., & Al Omari, H. (2013). Empirical investigation of the stability of IRT itemparameters estimation. International Online Journal of Educational Sciences, 5(2), 291–301. Disponível em: http://www.iojes.net//userfiles/Article/IOJES_980.pdf.

Azizan, N. H., Mahmud, Z., Rambli, A. (2020). Rasch Rating Scale Item Estimates using Maximum Likelihood Approach: Effects of Sample Size on the Accuracy and Bias of the Estimates. International Journal of Advanced Science and Technology Vol. 29, No. 4s, pp. 2526 – 2531.

Baker, F. B. (2001). The Basics of Item Response Theory. ERIC Clearinghouse on Assessment and Evaluation, College Park, MD. Second Edition. Disponível em: http://ericae.net/irt/baker.

Birnbaum, A. (1968). Some Latent Trait Models and Their Use in Inferring an Examinee's Ability. In: Lord, F.M. and Novick, M.R., Eds., Statistical Theories of Mental Test Scores.

Chuah, S. C., Drasgow F., & Luecht, R. (2006). How big is big enough? Sample size requirements for cast item parameter estimation. Applied Measurement in Education, 19(3), 241–255. Disponível em: http://dx.doi.org/10.1207/s15324818ame1903_5.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika. 16 (3): 297–334. doi:10.1007/BF02310555. hdl:10983/2196. S2CID 13820448. Disponível em: http://cda.psych.uiuc.edu/psychometrika_highly_cited_articles/cronbach_1951.pdf.

Linacre, J. M. (1994). Sample Size and Item Calibration [or Person Measure] Stability. Rasch Measurement Transactions, 1994, 7:4 p.328. Disponível em: https://www.rasch.org/rmt/rmt74m.htm.

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

Mallinson, T., Stelmack, J., Velozo, C. (2004). A Comparison of the Separation Ratio and Coefficient ⍁ in the Creation of Minimum Item Sets. Medical Care. Vol. 42, No. 1, Supplement: Applications of Rasch Analysis in Health Care, pp. 117–124. Disponível em: https://www.jstor.org/stable/4640698.

Novick, M. R.; Lewis, C. (1967). "Coefficient alpha and the reliability of composite measurements". Psychometrika. 32 (1): 1–13. doi:10.1007/BF02289400. PMID 5232569. S2CID 186226312.

Orlando M, Thissen D. Likelihood-based item-fit indices for dichotomous item response theory models. Appl Psychol Meas. 2000;24(1):50–64. Disponível em: https://www.researchgate.net/publication/236884113_Likelihood-Based_ItemFit_Indices_for_Dichotomous_Item_Response_Theory_Models.

Ree, M. J., & Jensen, H. E. (1980). Effects of sample size on linear equating of item characteristic curve parameters. In D. J. Weiss (Ed.), Proceedings of the 1979 Computerized Adaptive Testing Conference. Minneapolis, MN: University of Minnesota.

Şahin, A., & Anıl, D. (2017). The effects of test length and sample size on item parameters in item response theory. Educational Sciences: Theory & Practice, 17, 321–335. Disponível em: http://dx.doi.org/10.12738/estp.2017.1.0270.

Shojima, K., Toyoda, H. (2002). Estimation of Cronbach's alpha coefficient in the context of item response theory. School of Literature, YVaseda University, Toyama, Shinjuku-ku, Tokyo 162-8644. Disponível em: https://www.jstage.jst.go.jp/article/jjpsy1926/73/3/73_3_227/_pdf/-char/ja.

Sireci, S. G. (1991, June). Sample independent item parameters? An investigation of the stability of IRT item parameters estimated from small data sets. Paper presented at the annual Conference of Northeastern Educational Research Association, New York, NY.

Swaminathan, H., Hambleton, R. K., Sireci, S. G., Xing, D., & Rizavi, S. M. (2003). Small sample estimation in dichotomous item response models: Effect of priors based on judgmental information on the accuracy of item parameter estimates. Applied Psychological Measurement, 27(1), 27–51. Disponível em: http://dx.doi.org/10.1177/0146621602239475.

Tang, K. L., Way, W. D., & Carey, P. A. (1993). The effect of small calibration sample sizes on TEOFL IRT-based equating (TOEFL technical report TR-7). Educational Testing Service.

Robert L. Thorndike (1953, December). Who Belongs in the Family? Psychometrika. 18 (4): 267–276. Disponível em: https://doi.org/10.1007%2FBF02289263.

Expediente

Presidente
Carlos André

Diretores
Adriano Koelle, Andrés Kikuchi, Aquiles Mosca, Carlos Takahashi, César Mindof, Eduardo Azevedo, Eric Altafim, Fernanda Camargo, Fernando Rabello, Flavia Palacios, Giuliano De Marchi, Gustavo Pires, Julya Wellisch, Pedro Rudge, Roberto Paolino, Roberto Paris, Rodrigo Azevedo, Sergio Bini, Sergio Cutolo, Teodoro Lima e Zeca Doherty

Comitê Executivo
Amanda Brum, Eliana Marino, Francisco Vidinha, Guilherme Benaderet, Lina Yajima, Marcelo Billi, Soraya Alves, Tatiana Itikawa, Thiago Baptista e Zeca Doherty

Superintendente de Educação, Inovação e Sustentabilidade
Marcelo Billi

Gerente de Educação
Fernanda Mateus

Autoria
Andréa Mangabeira

Diagramação e Revisão
Fullbar

Contribuições
-

Rio de Janeiro
Praia de Botafogo, 501 - 704,
Bloco II, Botafogo,
Rio de Janeiro, RJ - CEP: 22250-911
Tel.: (21) 2104-9300

São Paulo
Av. Doutora Ruth Cardoso, 8501, 21°
andar, Pinheiros São Paulo, SP
CEP: 05425-070
Tel.: (11) 3471 4200


ANBIMA