ANÁLISE CLÁSSICA DOS ITENS A PARTIR DOS RESULTADOS RELATIVOS ÀS SUBESCALAS DE MATEMÁTICA "ESPAÇO E FORMA" E "GRANDEZAS E MEDIDAS" DA AVALIAÇÃO GLOBAL INTEGRADA – AGI

Cosme de Carvalho Rocha*

RESUMO

O presente trabalho aborda, dentro das teorias da Avaliação Educacional, a Teoria Clássica dos Testes e, de modo particular, sua aplicação na análise de dados reais do desempenho dos estudantes na Avaliação Global Integrada – AGI, realizada no primeiro semestre de 2019, tendo como público alvo a 3ª série do Ensino Médio Regular e Curso Técnico Integrado de todas as escolas da Rede Estadual de Educação do Piauí. Buscou-se analisar, a partir do desempenho dos estudantes, os itens da avaliação e, em particular, os relativos aos eixos ou subescalas de Matemática "Espaço e Forma e Grandezas e Medidas", realizando sua classificação quanto à dificuldade e discriminação, com o objetivo de verificar a qualidade desses itens e a adequação dos mesmos à população testada. Outro aspecto também analisado nesse estudo foi a confiabilidade do instrumento aplicado, com base na consistência interna dos seus itens, utilizando, para isso, o Coeficiente Alfa de Cronbach. A relevância desse estudo se justifica, entre outras razões, pelo fato de os indicadores e relatórios, gerados a partir da análise dos resultados dessa avaliação, subsidiarem as regionais de educação sobre o desenvolvimento das suas escolas, além de permitir acompanhar a implementação do currículo e, ainda, orientar a formação continuada dos professores, com foco na melhoria do desempenho dos estudantes nos componentes avaliados. Dentre os resultados alcançados, indica-se a possibilidade de uso, por coordenadores pedagógicos e professores, do conhecimento e qualidade dos itens das subescalas analisadas em relação ao nível de dificuldade e à sua capacidade discriminativa.

Palavras-chave: Avaliação educacional. Educação. Ensino. Teoria Clássica dos Testes. Rede Estadual de Ensino.

INTRODUÇÃO

Segundo Vilarinho (2015, apud KNÜPFER; AMARAL; HENNING, 2016), nas últimas décadas, tem crescido, de modo substancial, o uso das avaliações externas em larga escala para aferir o desempenho dos estudantes e, em assemelhada medida, também tem crescido a preocupação com a metodologia utilizada para obtenção desses resultados.

Isso ocorre, entre outras razões, porque a adoção de métodos e estatísticas inadequadas para a realização dessa aferição pode levar a resultados imprecisos e de baixa confiabilidade – o que poderia comprometer a leitura, análise e interpretação dos dados coletados, indicando desempenhos falsos dos educandos frente aos conteúdos abordados em sala de aula e esperados para o ano cursado.

Nessa perspectiva, este trabalho propõe abordar, dentro das teorias e métodos estatísticos da Avaliação Educacional, a Teoria Clássica dos Testes, e, de modo particular, aplicá-la na análise de dados reais do desempenho de 25.497 estudantes, em 487 escolas e 1138 turmas, da 3ª série do Ensino Médio da rede estadual de educação do Piauí.

O instrumento de avaliação usado para obtenção dos dados dessa pesquisa foi a Avaliação Global Integrada – AGI, realizada em 2019, e os dados foram fornecidos pela

* Coordenador de Avaliação Educacional e Currículo - CAEC, Secretaria Estadual de Educação do Piauí. E-mail: cosmecarvalho.rocha@gmail.com

Coordenação de Avaliação Educacional e Currículo (CAEC), da Secretaria de Estado da Educação do Piauí.

Esta avaliação é elaborada segundo uma Matriz de Referência baseada nas habilidades propostas no Currículo do Estado do Piauí, tem abrangência no 5º e 9º ano do Ensino Fundamental e na 3ª série do Ensino Médio e em todas as escolas da rede. A devolutiva pedagógica dos resultados é feita por meio de relatórios tratados em uma plataforma digital, o Mobieduca.Me/Mobicorretor.

Nesse contexto, o objetivo deste artigo consiste em analisar, a partir do desempenho dos estudantes na AGI, os itens dessa avaliação objetivando verificar sua qualidade. Para a realização dessa análise, fez-se um recorte nos microdados da avaliação, tratando apenas os itens relativos aos eixos ou subescalas de Matemática "Espaço e Forma e Grandezas e Medidas".

Desse modo, duas situações que se inter-relacionam mostraram-se importantes nessa fase do estudo: a primeira foi a necessária delimitação do escopo da pesquisa – por isso, a escolha de duas das subescalas da avaliação; e depois a classificação quanto à dificuldade e discriminação dos itens em análise. Tudo com o propósito, como dito, de verificar a qualidade desses itens e, consequente, adequação dos mesmos à população testada, uma vez que nenhum dos itens dessa avaliação haviam sido pré-testado, contribuindo, dessa forma, para o melhor entendimento do desempenho dos estudantes quando submetidos a esse tipo de avaliação.

É importante salientar que a Avaliação Global Integrada (AGI) – própria da rede estadual de educação do Piauí – é uma avaliação diagnóstica de caráter multidisciplinar, do tipo objetiva de múltipla escolha, aplicada trimestralmente aos estudantes do 5º e 9º ano do Ensino Fundamental e da 3ª série do Ensino Médio.

A elaboração desta avaliação é ancorada em uma Matriz de Referência baseada nas habilidades propostas no Currículo do Estado do Piauí relativas aos componentes curriculares de Língua Portuguesa e Matemática. Por outro lado, admitindo-se que o nível de qualidade do ensino é determinante para o sucesso dos estudantes, a aferição desse desempenho se torna uma importante ferramenta para se verificar essa qualidade.

Assim sendo, é de grande valia qualificar, no âmbito da Secretaria de Estado da Educação do Piauí, o debate em torno dessa avaliação, oportunizando às equipes escolares uma apropriação independente e objetiva das devolutivas pedagógicas dessas avaliações objetivando melhorar o nível de qualidade do ensino.

Outro aspecto igualmente importante, também analisado nesse estudo, é a confiabilidade do instrumento aplicado, com base na consistência interna dos seus itens, utilizando, para isso, o Coeficiente Alfa de Cronbach.

Ressalta-se que este estudo está respaldado em autores como: Silveira (1983); Pasquali (1996); Freitas e Rodrigues (2005); Erthal (2009); Borgatto e Andrade (2012); Vilarinho (2015), entre outros.

Borgatto e Andrade (2012), citados por Knüpfer, Amaral e Henning (2016), por exemplo, afirmam que a avaliação do desempenho dos estudantes depende, fundamentalmente, da qualidade dos itens da prova. Nesse sentido, a prova como instrumento de avaliação, deve ser bem construída e os itens que a compõem precisam ter as propriedades psicométricas que os qualifiquem como confiáveis.

Partindo dessa premissa, é recomendável verificar as características desses itens, na medida em que itens muito difíceis ou muito fáceis acrescentam pouco à confiabilidade da prova, sendo necessário que sejam substituídos.

Pasquali (1996), por sua vez, ao estudar a análise de itens, constata a existência de dois modelos distintos: um identificado como Análise Teórica; e outro, como Análise Empírica ou Estatística. O primeiro é feito por juízes e busca estabelecer a compreensão dos itens e a

pertinência dos mesmos ao atributo que pretendem medir. Já o segundo, a Análise Empírica ou Estatística, prescreve que as propriedades psicométricas de cada item em um teste são dadas pelos seguintes parâmetros: Índice de Dificuldade e Índice de Discriminação.

Quanto à relevância acadêmica e social desse estudo, este se justifica, entre outras razões, pelo fato de os indicadores e relatórios gerados a partir da análise dos resultados dessa avaliação subsidiarem as regionais de educação sobre o desenvolvimento das suas escolas, além de permitir acompanhar a implementação do currículo – verificando se os professores estão realmente aplicando o currículo proposto e se os alunos estão efetivamente aprendendo – e, por último, e igualmente importante, orientar a formação continuada dos professores com foco na melhoria progressiva do desempenho dos estudantes nos componentes avaliados.

Finalmente, por zelo metodológico, este artigo encontra-se estruturado em cincos seções, incluindo essa introdução. A revisão da literatura sobre a Teoria Clássica dos Testes - TCT na 2ª seção; na terceira seção, é abordada a Avaliação Global Integrada – AGI e os procedimentos metodológicos para análise dos itens propostos referente às subescalas "Espaço e Forma" e "Grandezas e Medidas"; na 4ª seção, é apresentada a análise dos resultados referentes à dificuldade e discriminação dos itens, além da confiabilidade da avaliação e, finalmente, na última seção, são apresentadas as considerações finais, e as referências bibliográficas.

2 A REVISÃO DE LITERATURA/TEORIA CLÁSICA DOS TESTES-TCT

Nessa seção, fez-se uma breve revisão da Teoria Clássica dos testes, abordada nos livros e artigos que constam da bibliografia, com foco no Índice de Dificuldade, Índice de Discriminação, nos coeficientes de correlação ponto bisserial (Pbp) e Índice D, além das Medidas de confiabilidade de um teste através do Coeficiente alfa de Cronbach.

2.1 Teoria Clássica dos testes

Segundo Piton-Gonçalves e Almeida (2018), a elaboração e análise de um teste pode seguir a Teoria Clássica de Testes (TCT) (ERTHAL, 2001; PASQUALI, 2009) e/ou a Teoria de Resposta ao Item (TRI) (LORD, 1980; ANDRADE; TAVARES; VALLE, 2000). É relevante ressaltar que ambos os métodos são complementares, no caso desse estudo, todas as análises são fundamentadas na TCT.

Referenciada nessa teoria, a avaliação do desempenho dos estudantes está intrinsecamente relacionada a três variáveis, igualmente importantes, quais sejam: a qualidade dos itens que compõe a prova; o tamanho do teste (número de itens); e por fim, os indivíduos que a fizeram. Aqui, é valido o raciocínio de que todas as interpretações estão associadas à avaliação e à coorte que a respondeu.

Notadamente, no cenário educacional brasileiro, é prática corrente avaliar o desempenho dos estudantes por meio de testes construídos com itens objetivos de múltipla escolha. Nesse tipo de avaliação, a TCT se configura como método mais utilizado para a sua correção, isso se deve ao fato de o escore total do teste se basear na quantidade total de itens respondidos corretamente pelo examinando. Um dos objetivos da TCT é interpretar o significado desse escore, verificando, desta forma, a qualidade do teste aplicado.

Nesse contexto, de acordo com Andrade e Borgatto (2012, apud KNÜPFER; AMARAL; HENNING, 2016) e Pasquali (1996), as propriedades psicométricas dos itens de um teste avaliativo tem estreita relação com os seguintes parâmetros: índice de dificuldade, índice de discriminação e correlação bisserial, cujo detalhamento é apresentado a seguir.

2.1.1 Índice de Dificuldade (ID)

A Teoria Clássica dos Testes – TCT – configura-se como o modelo mais conhecido e adotado pelos professores para se avaliar o desempenho dos estudantes. Nesse modelo, o resultado de uma prova leva em consideração apenas os escores brutos dos indivíduos que a realizaram. Desse modo, o resultado dela é obtido verificando-se a quantidade de itens respondidos corretamente por cada respondente.

Ainda com base nesse modelo, os escores totais dos estudantes obtidos na prova dependem do teste utilizado. Logo, o desempenho do estudante em um determinado teste pode variar em função da facilidade ou dificuldade de seus itens. Desta forma, a lógica nos impõe o seguinte raciocínio: quando um teste é difícil, o estudante tende a apresentar baixo desempenho (escore total baixo) e, quando é fácil, tende a apresentar alto desempenho (escore total alto).

Para Pasquali (1996, p. 82), "a dificuldade do item é definida em termos da porcentagem (proporção) de sujeitos que dão respostas corretas (testes de aptidão) ou de acordo/preferência (testes de personalidade) ao item".

Diante do exposto, é pertinente e recomendável conhecer o Índice de Dificuldade (ID) dos itens de um teste, pois, assim, é possível descartar itens que possuem um nível de dificuldade bastante elevado, assim como os itens que possuem dificuldade muito baixa, isso se explica pelo fato de que itens dessa natureza nada ou quase nada têm a nos informar sobre as diferenças individuais dos respondentes.

O Índice de Dificuldade (ID) propõe-se, portanto, a medir as diferenças individuais no que diz respeito ao rendimento alcançado no teste. Esse índice, específico para cada item, pode ser calculado estabelecendo-se a razão entre o número de indivíduos que acertaram e o número total de indivíduos submetidos a esse item em particular. Igualmente, quanto menor a porcentagem de acerto maior será o grau de dificuldade, ou seja, quanto mais sujeitos erram determinado item, mais difícil ele é.

O Índice de Dificuldade varia de 0 a 1, ($0 <= ID <= 1$), e quanto mais próximo de 1 (um), mais fácil é o item. Erthal (2009) o denomina de Índice de Facilidade, pois quanto mais sujeitos acertam o item, mais fácil ele é.

O valor do Índice de Dificuldade (ID) está diretamente relacionado à média do teste e pode ser calculado pela fórmula:

$$ID_i = \frac{A}{N} \qquad (1)$$

Onde:

ID_i: É o índice de dificuldade
A: É o número de acertos no teste
N: É o número de indivíduos que responderam ao teste.

Conforme Pasquali (2013, apud NAGEL, 2018, p. 36), "para que uma avaliação educacional tenha nível de dificuldade ideal, é preciso que seus índices de dificuldades estejam distribuídos em conformidade com uma curva normal", como mostra a tabela 1, a seguir:

Tabela 1 - Classificação e percentuais esperados para os Índices de Dificuldade na TCT

Intervalo do ID	Percentual	Classificação do item
ID > 0,9	10%	Muito Fácil
0,7 ≤ ID ≤ 0,9	20%	Fácil
0,3 ≤ ID ≤ 0,7	40%	Normal (ou Moderado)
0,1 ≤ ID < 0,3	20%	Difícil
ID < 0,1	10%	Muito Difícil

Fonte: Adaptado de Pasquali (2013).

Vale ressaltar que essa classificação e percentuais esperados para os Índices de Dificuldade foram utilizadas como critério para a análise do Índice de Dificuldade ID dos itens da AGI. Não obstante, nas subescalas analisadas, essa distribuição não necessariamente precisa seguir uma distribuição normal.

2.1.2 Índice de Discriminação (IDCS)

Como já mencionado ao longo desse trabalho, algumas variáveis interferem diretamente no desempenho dos estudantes em uma prova, dente elas, destacamos a qualidade dos itens que a compõem e a coorte que a respondeu.

Assim, é razoável admitir que, numa prova, como a Avaliação Global Integrada (AGI), que contou com a participação 25.497 estudantes, alguns itens sejam considerados fáceis pelos estudantes que obtiveram alto desempenho na prova (escore total alto), e outros possam ser considerados difíceis pelos que obtiveram baixo desempenho (escore total baixo). Notadamente isso se deve em virtude de, na TCT, os parâmetros dos itens serem dependentes do grau de domínio dos conteúdos por parte dos sujeitos que fizeram a prova.

Nessa perspectiva, com base em Pasquali (2009) e Viana (1976), Sousa (2018, p. 32) aponta que "a discriminação na TCT é definida como a capacidade do item distinguir sujeitos de escores altos em relação àqueles de escores baixos ou diferenciar sujeitos de desempenho baixo e superior".

Para Erthal (2009), o Índice de Discriminação estabelece relação entre escores totais altos ou baixos no teste e as respostas corretas ou incorretas dadas a um item. Desse modo, de acordo com Gomes (2014, p. 18), "o parâmetro de discriminação é o responsável por aumentar, ou não, a diferença entre as probabilidades de alunos com desempenhos distintos responderem corretamente o item".

Vale salientar que Erthal (2009, p. 82) considera: "um erro enorme construir um teste selecionando itens de acordo apenas com a sua dificuldade. Essa informação é importante, mas secundária em relação àquela obtida pela intercorrelação entre os itens". Portanto, itens considerados bons devem ter uma correlação discreta com outros itens e uma correlação elevada com o escore total do teste.

Ainda, segundo Erthal (2009 p. 82), "na elaboração de um teste, o que se procura são itens com pequeno ou moderado grau de inter-relacionamento, mas com boa correlação com o escore total. Essas informações são obtidas pelo Índice de Discriminação".

Na psicometria tradicional, existe um número variado de técnicas estatísticas para estabelecer o Índice de Discriminação (IDS) de um item. No entanto, Erthal (2009, p. 83) afirma que se pode calcular o IDS pela correlação bisserial por pontos pela correlação bisserial simples e pelo método dos 27%. Para efeito deste estudo, nossa atenção voltou-se

especialmente para dois desses métodos: o índice D e a correlação ponto bisserial ou polisserial.

O índice D, por ser de cálculo mais simples e de fácil entendimento; e a correlação ponto bisserial, por ser a mais adequada quando se trata de testes de rendimento escolar – objeto desse artigo –, além de ser menos influenciada pela dificuldade do item.

2.1.2.1 índice D

Para Erthal (2009, p. 81), "o Índice de Discriminação estabelece relação entre escores totais altos ou baixos no teste e as respostas corretas ou incorretas dadas a um item". Desse modo, a divisão dos indivíduos é feita distribuindo-os em dois grupos: (i) grupo superior (ACIM) - são os indivíduos que estão enquadrados nos 27% que obtiveram os escores mais altos; e (ii) grupo inferior (ABAI) - os que se acham na outra extremidade que obtiveram, portanto, os escores mais baixos. O valor da diferença entre o número de acertos nesses dois grupos é denominado de índice D e, conforme a literatura, varia de 0 a 100 se for calculado em porcentagem ou de 0 a 1, desde que em proporção.

Como o grupo superior (ACIM) reúne os respondentes com maior desempenho no teste, a lógica nos impõe que a proporção de acertos para um item em particular seja maior nesse grupo do que no grupo inferior (ABAI). Dessa maneira, quanto maior a diferença entre os percentuais de acertos nesses dois grupos, maior será a discriminação do item.

Portanto, segundo Erthal (2009, p. 86) "o Índice Discriminação D (IDS) é um parâmetro bastante razoável para indicar se um item é capaz de diferenciar os participantes que obtiveram alto ou baixo desempenho na prova", e segundo Erthal (2009, p. 86), "pode ser calculado pela fórmula":

$$IDS_D = \frac{A_s - A_i}{\frac{n_s + n_i}{2}} \qquad (2)$$

onde:
IDS_D = Índice de Discriminação D
A_s = acertos do grupo superior (27%)
A_i = acertos do grupo inferior (27%)
n_s = número de indivíduos que atingiram o grupo superior
n_i = número de indivíduos que atingiram o grupo inferior

É importante destacar, também, que uma classificação bastante aceita para a discriminação dos itens de um teste é a proposta por Ebel (1954, apud PITON-GONÇALVES e ALMEIDA, 2018), mostrada na Tabela a seguir e que foi utilizada na análise do Índice de Discriminação D da AGI (Tabela 2):

Tabela 2 - Classificação do Índice de Discriminação

Intervalo do IDS – D	Classificação do item
Discriminação < 0,20	Ineficienete (deve ser rejeitado)
0,20 ≤ Discriminação < 0,30	Item Marginal (necessita revisão)
0,30 ≤ Discriminação < 0,40	Aceitável (sujeito aprimoramento)
Discriminação ≥ 0,40	Adequado, devendo permanecer no teste.

Fonte: Adaptada de Ebel (1954).

Pela interpretação dos dados da tabela 2, observa-se que os itens que devem permanecer no teste são os classificados como aceitáveis ou satisfatórios. Na mesma direção e sentido, itens que necessitam de revisão deverão passar por ajustes para serem reincluídos. No entanto, aqueles classificados como ineficientes recomenda-se que sejam eliminados do teste.

Desse modo, espera-se que, em uma avaliação educacional, o poder de discriminação do item seja superior ou igual a 0,4 conforme a tabela 2. No entanto, ainda de acordo com a tabela 2, para esse estudo, foram considerados discriminativos os itens com índice D ≥ 0,30.

2.1.2.2 Coeficiente de Correlação Ponto Bisserial

O Coeficiente de correlação ponto bisserial é uma medida estatística que mede (como o nome o próprio nome já diz) a correlação entre o desempenho do respondente no item e no teste como um todo, sendo, portanto, uma medida da capacidade de discriminação do item em relação ao resultado do teste. Entretanto, é mais sofisticado do que o índice D e menos influenciado pela dificuldade do item, além de ser, conforme a literatura, a mais adequada quando se trata de testes de rendimento escolar, como é o caso desse artigo.

A Correlação Ponto Bisserial é, em certa medida, uma versão resumida do Coeficiente de Correlação de Pearson, que se utiliza quando uma variável é contínua (pontuação total no teste) e a outra, nesse caso o item, é dicotômica discreta, ou seja, quando só há duas possibilidades de resposta (pontuação de acerto ou erro no item). Assim, ao analisar os itens, eles são dicotomizados, em que 0 (zero) corresponde a errado e 1 a certo.

Da mesma forma que a correlação de Pearson varia entre -1 e 1, a estatística desse coeficiente de correlação deve indicar uma relação direta, ou seja, sua variação fica definida entre os valores de 0 a 1. Para esse estudo, foram considerados discriminativos os itens com correlação ponto bisserial (ρpb) ≥ 0,30.

Como observado, essa correlação é indicada quando a distribuição dos dados é dicotômica, assimétrica, e conforme Erthal (2009, p. 85), "o modelo matemático dessa técnica é dado da seguinte forma":

$$\rho_{pb} = \frac{\overline{X}_A - \overline{X}_T}{S_t} \sqrt{\frac{p}{1-p}} \qquad (3)$$

em que,

ρ_{pb} = correlação ponto bisserial;
$\overline{X}_A$ = média dos sujeitos que acertam o item no teste;
$\overline{X}_T$ = média total do teste;
S_t = desvio padrão do teste;
p = proporção de sujeitos que acertaram o item;

Valores altos para a correlação ponto bisserial nos indicam que esses itens são mais capazes de separar examinandos com níveis semelhantes de desempenho. Não obstante, para Erthal (2009, p. 82), itens "ruins apresentam uma correlação nula e uma correlação negativa indica-nos que o indivíduo se saiu bem no teste, embora tenha tido um mau desempenho no item".

De acordo com Vilarinho (2015, apud NAGEL, 2018, p. 37), "espera-se que o gabarito do item (questão correta), apresente correlação positiva, e que seus distratores (questões

erradas), valores negativos. Desta forma, podemos afirmar que os alunos de melhor desempenho no teste, estão acertando o item".

2.1.3 Confiabilidade de um teste

Para Soares (2018), a primeira análise a ser feita em um teste, com base na Teoria Clássica dos Testes, diz respeito à sua confiabilidade ou consistência interna, para verificar a qualidade do instrumento de medida na mensuração das habilidades dos respondentes. A confiabilidade ou fidedignidade trata, portanto, da estabilidade dos resultados e é desejável que eles sejam os mais consistentes possíveis.

Segundo Gaspar e Shimoya (2017), "O coeficiente alfa de Cronbach, descrito por Lee J. Cronbach (CRONBACH, 1951), é uma das ferramentas estatísticas mais importantes e difundidas em pesquisas que envolvem a construção de testes e sua aplicação".

De acordo com Almeida, Santos e Costa (2010, apud NAGEL, 2018, p. 38), o coeficiente alfa pode ser conceituado como "a medida pela qual algum constructo, conceito ou fator medido está presente em cada item".

O referido coeficiente "mede a correlação entre respostas em um questionário através da análise do perfil das respostas dadas pelos respondentes" (HORA et al., 2010, apud MATTHIENSEN, 2011, p. 12). Partindo do raciocínio de que todos os itens de uma prova utilizam a mesma escala de medição, Matthiensen (2011, idem) enfatiza que o cálculo do coeficiente α se dá "a partir do somatório da variância dos itens individuais e da soma da variância de cada avaliando", pela expressão a seguir:

$$\alpha = \left(\frac{K}{K-1}\right) \mathrm{x} \left(1 - \frac{\sum S_i^2}{S_T^2}\right) \qquad (4)$$

Onde,

K é o numero de itens;
S_i^2 é a soma das variâncias dos n itens;
S_T^2 é a variância global dos escores dos testes.

De acordo com a literatura, a confiabilidade do Coeficiente alfa de Cronbach normalmente varia entre 0 e 1. Quanto mais próximo de 1 ele estiver, melhor será sua precisão, caracterizando, assim, uma segurança para a medida do fenômeno que se quer avaliar.

Segundo Hora, Monteiro e Arica (2010), apesar de existir na literatura científica uma ampla e abrangente aplicação do Coeficiente Alfa de Cronbach nas diversas áreas do conhecimento, ainda não existe um consenso entre os pesquisadores acerca da interpretação da confiabilidade de um teste obtida a partir do valor deste coeficiente.

No entanto, de acordo com Almeida, Santos e Costa (2010, apud NAGEL, 2018, p. 38), o valor mínimo aceitável em avaliações educacionais para o alfa é 0,70. Números abaixo desse valor indicam um instrumento inconsistente e pouco confiável.

Para análise da consistência interna, com o uso do Coeficiente Alfa de Cronbach, utilizamos, nesse estudo, a classificação sugerida por Freitas e Rodrigues (2005, apud GASPAR e SHIMOYA, 2017), cujos limites estão apresentados na tabela 3, a seguir.

Tabela 3 - Valores do coeficiente α de Cronbach

Valor de α	Confiabilidade do Teste
$\alpha \leq 0{,}30$	Muito baixa
$0{,}30 < \alpha \leq 0{,}60$	Baixa
$0{,}60 < \alpha \leq 0{,}75$	Moderada
$0{,}75 < \alpha \leq 0{,}90$	Alta
$\alpha > 0{,}90$	Muito Alta

Fonte: Elaboração própria com base nos dados de Freitas e Rodrigues (2005).

No entanto, como a decisão a respeito do valor mínimo de confiabilidade de um teste fica a critério do pesquisador, para esse estudo, foram considerados adequados os valores de $\alpha \geq 0{,}7$ sugeridos por Almeida, Santos e Costa (2010).

3 AVALIAÇÃO GLOBAL INTEGRADA – AGI

A Avaliação Global Integrada – AGI, objeto de análise nesse estudo, é um dos instrumentos permanentes de avaliação e acompanhamento tanto do currículo como do ensino e aprendizagem da rede estadual de ensino do Piauí.

Criada em 2018, pela Coordenação de Avaliação Educacional e Currículo (CAEC), da Secretaria de Estado da Educação do Piauí, a AGI segue uma Matriz de referência baseada nas habilidades propostas no Currículo da rede estadual, tendo como unidades de interesse: verificar se o currículo proposto está sendo aplicado; fornecer subsídios para que professores e gestores identifiquem o que os alunos estão aprendendo e, por último, mas não menos importante, orientar um plano de intervenção pedagógico para corrigir possíveis defasagens de aprendizagem encontradas.

Embora sendo obrigatória, sua nota não entra na composição da media bimestral do estudante nos níveis de ensino, anos/séries, disciplinas e bimestres definidos pela SEDUC/PI. É uma avaliação diagnóstica de caráter multidisciplinar, do tipo objetiva de múltipla escolha, aplicada trimestralmente a todos os estudantes devidamente matriculados no 5º e 9º ano do Ensino Fundamental e na 3ª série do Ensino Médio. O estudante tem seu desempenho apurado numa escala de notas variando de 0 (zero) a 10 (dez).

Como se pode observar, essa avaliação configura-se como um instrumento valiosíssimo para os professores, ao passo que subsidia suas ações para a tomada de decisões mais assertivas sobre sua prática pedagógica e orientação de suas estratégias de atuação.

3.1 Procedimento de analise da Avaliação Global Integrada - AGI

Buscando aplicar a TCT no tratamento dos dados de desempenho dos estudantes na Avaliação Global Integrada - AGI, com o objetivo de estimar os índices de dificuldade e discriminação, além da análise da consistência interna dos itens dessa avaliação como um todo, foi feita uma análise inicial dos dados relativos aos resultados da primeira edição da avaliação realizada em junho de 2019. Tais dados foram obtidos por meio de uma plataforma digital: o MobiEduca.ME/Mobicorretor, disponibilizado pela Secretaria de Estado da Educação do Piauí – SEDUC/PI e, a partir daí, realizou-se a seleção dos itens de matemática relativos às subescalas: Espaço e Forma, e Grandezas e Medidas; nas quais se centra esta produção científica.

3.1.1 Metodologia

Para esse estudo, utilizaram-se dados disponibilizados pela Coordenação de Avaliação Educacional e Currículo – CAEC da SEDUC/PI, relativos à primeira edição da Avaliação Global Integrada – AGI, realizada em junho de 2019. Primeiro foi feita a seleção dos itens pertencentes aos eixos do conhecimento ou subescalas de Matemática relativas à "Espaço e Forma; e Grandezas e Medidas" num total de 7 itens. A análise subsequente se concentrou nos parâmetros de dificuldade e discriminação dos itens de cada uma dessas subescalas. A confiabilidade da avaliação, como um todo, também foi estimada, utilizando para este fim o Coeficiente Alfa de Cronbach.

Nessa edição, foram analisadas as respostas de estudantes da 3ª série do Ensino Médio Regular de todas as escolas da Rede Estadual de Ensino e também das escolas de Ensino Médio Integrado - Curso Técnico Integrado. A base de dados contém informações de 25.497 estudantes, em 487 escolas e 1138 turmas, envolvidas no levantamento.

3.1.2 Procedimentos para análise dos dados

Inicialmente, realizou-se um estudo exploratório dos dados. Os itens analisados tiveram por base as pontuações obtidas na prova e foram analisados segundo os parâmetros da dificuldade e discriminação. Tanto o índice de dificuldade quanto o de discriminação da referida avaliação foram estimados com base na Teoria Clássica dos testes – TCT.

Para o estudo da dificuldade, estimada pela equação (1), foi considerada a proporção de respostas certas ao item. Para se estimar a discriminação, foi considerado o Índice D, calculado pela equação (2) e a Correlação Ponto Bisserial; estimada pela equação (3). Quanto à confiabilidade do teste, essa foi estimada com base na consistência interna dos seus itens, utilizando para este fim o Coeficiente Alfa de Cronbach mensurado de acordo com a equação (4).

3.1.2.1 Resultado1: Classificação dos conteúdos

Os conteúdos matemáticos foram classificados de acordo com os Domínios e Competências que sistematizam a Matriz de Referência de Matemática do Ensino Médio em quatro eixos temáticos ou subescalas: Espaço e Forma (eixo I); Grandezas e Medidas (eixo II); Números e Operações/Álgebra e Funções (eixo III) e Tratamento da Informação (eixo IV). Dessa forma, associaram-se os conteúdos da Matriz Curricular do Ensino Médio, culminando com o que se denominou de Subescalas, sumarizadas na Tabela 4.

Tabela 4 - Eixos de Conhecimento da AGI da 3ª série Ensino Médio

Eixos ou Subescalas	Numeração dos itens	Representatividade
(I) – Espaço e Forma	01, 08, 14, 16	15,4%
(II) – Grandezas e Medida	02, 18, 22,	11,5%
(III) – Números e Operações/Álfgebra e Funções	03, 04, 05, 06, 07, 10, 11, 12, 13, 15, 17, 19, 20, 21, 23, 26 e 24	65,4%
(IV) – Tratamento da Informação	09, 25	7,7%
		100%

Fonte: Produção própria (2020).

3.1.2.2 Resultado 2: Análise Clássica dos Itens (1ª AGI/2019)

As próximas seções trazem as análises dos itens de Matemática, relativos às subescalas "Espaço e Forma, e Grandezas e Medidas" da 1ª AGI realizada em junho de 2019, selecionados para este artigo, além da análise da confiabilidade do instrumento como um todo.

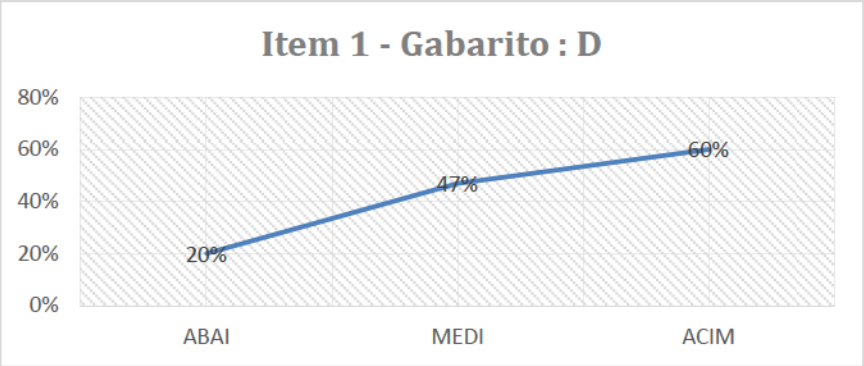
Neste tópico, optou-se por utilizar a seguinte dinâmica, para melhor compreensão dos dados apresentados: será apresentada uma tabela com as características psicométricas clássicas de cada item e, logo em seguida, um gráfico com o padrão de resposta do gabarito referente ao item analisado, Após essa apresentação, tem-se a análise do item, já com base na tabela tanto quanto no gráfico.

Tabela 5 - Características psicométricas clássicas do Item 1

Estatisticas Clássicas															
ITEM 1															
ÍNDICES					PERCENTUAIS DE RESPOSTAS					COEFICIENTES BISSERIAIS				
ID	DISCR	ABAI	ACIM	BISE	A	B	C	D	E	A	B	C	D	E
0.39	0.43	0.20	0.60	0.38	30.9	8.2	13.6	38.7	7.9	-0,26	-0,18	-0,05	0,38	-0,12

Fonte: Produção própria, 2020

Gráfico 1. Análise do Item 1: Frequência de escolha do gabarito nos três grupos de desempenho



Fonte: Produção própria (2020).

Os grupos referidos neste item são: Grupo 1- (ABAI) 27% de menor desempenho, Grupo 2- (MEDI) 46% com desempenho mediano e Grupo 3- (ACIM) 27% de maior desempenho.

O item 1 é do eixo de conhecimento ou subescala "Espaço e Forma" e o Descritor (habilidade) da matriz de referência da avaliação cobrado nele foi o D5 "*resolver problema que envolva razões trigonométricas no triângulo retângulo (seno, cosseno, tangente)*". Pela análise, o item obteve índice de dificuldade (ID) = 0,39 sendo classificado como mediano ou moderado.

A partir do gráfico 1, com o padrão resposta do gabarito nos três grupos de desempenho, percebe-se que 20% dos participantes que pertencem ao grupo de menor desempenho (ABAI), acertaram o item, enquanto que, no grupo de maior desempenho (ACIM), esse percentual foi de 60%, significando que a maioria dos alunos de alto desempenho no teste acertaram esse item em particular.

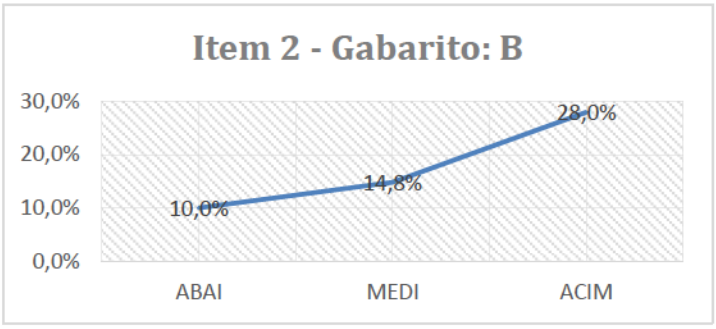
Desse modo, o *índice de discriminação D* apurado foi igual a 0,43 – o que indica ser um item satisfatório. Em relação aos coeficientes ponto bisseriais, o gabarito (D) apresentou valor positivo (0,38), com aceitável poder de discriminação. As outras alternativas (distratores) apresentaram valores negativos, como esperado, e, de acordo com Nagel (2018), indicando que os alunos de alto desempenho na prova escolheram menos estas alternativas do que os alunos de baixo desempenho.

Tabela 6 - Características psicométricas do Item 2

Estatisticas Clássicas																
ITEM 2																
ÍNDICES					PERCENTUAIS DE RESPOSTAS					COEFICIENTES BISSERIAIS						
ID	DISCR	ABAI	ACIM	BISE	A	B	C	D	E	A	B	C	D	E		
0.17	0.18	0.10	0.28	0.23	28.4	17.3	28.4	20.4	8.3	-0,03	0,23	-0,06	-0,04	-0,07		

Fonte: Produção própria, 2020

Gráfico 2. Análise do Item 2: Frequência de escolha do gabarito nos três grupos de desempenho (ABAI) (MEDI) e (ACIM)



Fonte: Produção própria (2020).

Este item pertence ao eixo do conhecimento ou subescala "Grandezas e Medidas" e o Descritor da matriz de referência da avaliação cobrado foi o D13 "*resolver problema envolvendo a área total e/ou volume de um sólido (prisma, pirâmide, cilindro, cone, esfera*)". Obteve-se, nesse item, *índice de dificuldade* (ID) = 0,17 (teve um baixo índice de acerto, apenas 17%), sendo classificado como difícil.

A partir do gráfico 2, relativo a esse item, nota-se que apenas 10% dos participantes pertencentes ao grupo de menor desempenho (ABAI) acertaram o item em relação ao grupo de maior desempenho (ACIM). Esse percentual também foi relativamente baixo, dos melhores alunos no teste apena 28% acertaram esse item em particular. Desse modo, o item apresentou *Índice de Discriminação D* igual a 0,18, que indica ser um item Ineficiente. Portanto, conforme a classificação sugerida por Ebel (1954) deveria ser rejeitado.

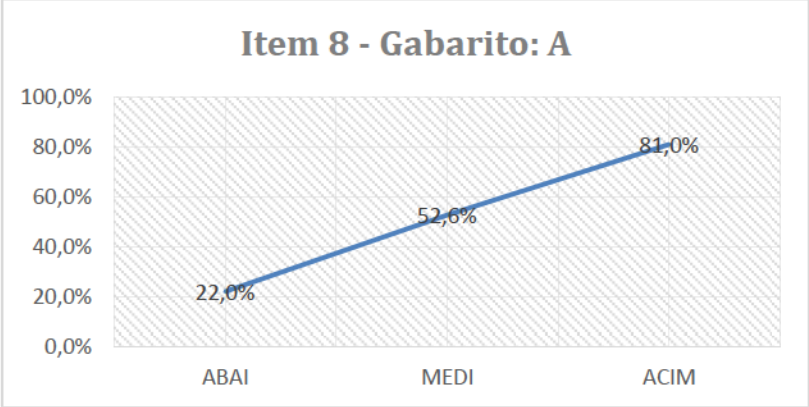
Em relação aos coeficientes correlação ponto bisseriais, o gabarito (B) apresentou valor positivo (0,23) com baixo poder de discriminação. Já os distratores, como esperado, apresentaram valores negativos, indicando serem as opções mais escolhida pelos os alunos de baixo desempenho na prova.

Tabela 7 - Características psicométricas do Item 8

Estatisticas Clássicas														
ITEM 8														
ÍNDICES					PERCENTUAIS DE RESPOSTAS					COEFICIENTES BISSERIAIS				
ID	DISCR	ABAI	ACIM	BISE	A	B	C	D	E	A	B	C	D	E
0.52	0.59	0.22	0.81	0.44	30.7	8.7	7.0	51.7	2.0	-0,30	-0,23	-0,29	-0,44	-0,12

Fonte: Produção própria, 2020

Gráfico 3. Item 8: Frequência de escolha do gabarito nos três grupos de desempenho (ABAI) (MEDI) e (ACIM)



Fonte: Produção própria (2020).

O item 8 abordou o Descritor de número 6 da matriz de referencia, qual seja: "*identificar a equação de uma reta apresentada a partir de dois pontos dados ou de um ponto e sua inclinação*)". É da subescala "Espaço e Forma" e, de acordo com os resultados, apresentou *índice de dificuldade* (ID) = 0,52 sendo classificado como moderado ou mediano visto que 52% dos respondentes o acertaram.

Quanto ao Índice de Discriminação, analisando o gráfico 3, visualiza-se que 22% dos acertos está no grupo de menor desempenho (ABAI), enquanto que no grupo de maior desempenho (ACIM), esse percentual é alto 81% - número bastante considerável, significando que 81% dos melhores alunos no teste acertaram esse item em particular. Desse modo, o *Índice de Discriminação D*, estabelecido pela diferença entre esses dois percentuais tem valor igual a 0,59 indicando um item adequado para a avaliação, pois conseguiu distinguir bem os alunos com desempenho distintos na prova.

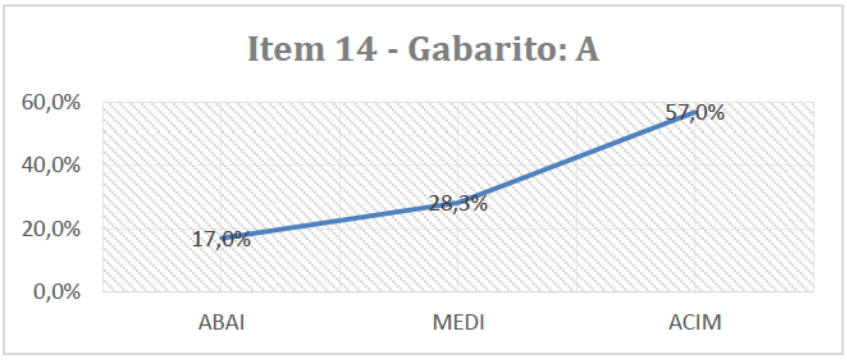
Em relação aos coeficientes de correlação ponto bisseriais, o gabarito (D) apresentou valor positivo (0,44) com elevado poder de discriminação, todos os distratores apresentaram bisserial negativo indicando que os alunos com bom desempenho, não procuraram escolher essas alternativas.

Tabela 8 - Características psicométricas do Item 14

Estatisticas Clássicas														
ITEM 14														
ÍNDICES					PERCENTUAIS DE RESPOSTAS					COEFICIENTES BISSERIAIS				
ID	DISCR	ABAI	ACIM	BISE	A	B	C	D	E	A	B	C	D	E
0.33	0.40	0.17	0.57	0.36	33,4	26,4	24,5	13,3	2,6	0,36	-0,13	-0,14	-0,19	-0,22

Fonte: Produção própria, 2020

Gráfico 4. Análise do Item 14: Frequência de escolha do gabarito nos três grupos de desempenho (ABAI) (MEDI) e (ACIM)



Fonte: Produção própria (2020).

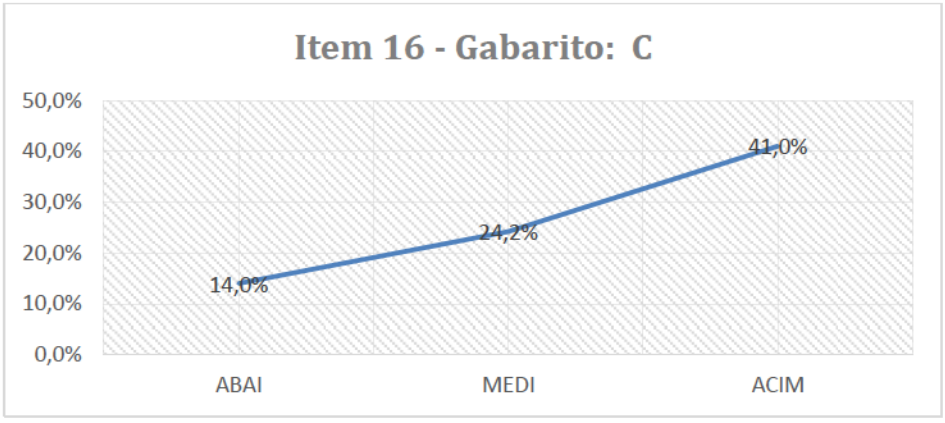
O item 14, pertencente à subescala "Espaço e Forma" e o Descritor da matriz de referência da avaliação cobrado foi o D5 *"resolver problema que envolva razões trigonométricas no triângulo retângulo (seno, cosseno, tangente)*". Obteve-se *índice de dificuldade (ID)* = 0,33 sendo classificado como moderado mediano, e *Índice de Discriminação D* igual a 0,40 que indica ser um bom item, pois conseguiu distinguir os alunos com baixo e alto desempenho. Em relação aos coeficientes correlação ponto bisseriais, o gabarito (A) apresentou valor positivo (0,36), com aceitável poder de discriminação para os padrões da avaliação. As alternativas (B), (C) (D) e (E) apresentaram valores negativos, como esperado, indicando que essas alternativas foram mais procuradas pelos alunos com baixo desempenho.

Tabela 9 - Características psicométricas do Item 16

Estatisticas Clássicas														
ITEM 16														
ÍNDICES					PERCENTUAIS DE RESPOSTAS					COEFICIENTES BISSERIAIS				
ID	DISCR	ABAI	ACIM	BISE	A	B	C	D	E	A	B	C	D	E
0.26	0.27	0.14	0.41	0.30	27,7	27,0	18,1	16,7	12,4	0,30	-0,10	0,10	-0,13	-0,16

Fonte: Produção própria, 2020

Gráfico 5. Análise do Item 16: Frequência de escolha do gabarito nos três grupos de desempenho (ABAI) (MEDI) e (ACIM)



Fonte: Produção própria (2020).

No item 16, pertencente ao eixo de conhecimento ou subescala "Espaço e Forma", o Descritor cobrado foi o D1 - Identificar figuras semelhantes mediante o reconhecimento de relações de proporcionalidade. Obteve-se ID = 0,26 sendo classificado como difícil, e *Índice D* igual a 0,27 que indica ser, de acordo com a classificação de Ebel (1956), um item marginal, portanto necessita revisão.

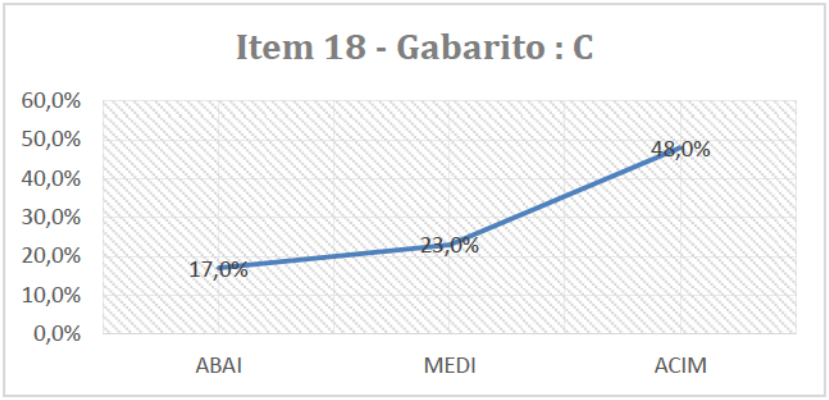
Em relação aos coeficientes correlação ponto bisseriais, o gabarito (A) como esperado, apresentou valor positivo (0,30) com moderado poder de discriminação. As alternativas (B), (D) e (E) apresentaram valores negativos, significando que os alunos de alto desempenho no teste não escolheram estas alternativas. Não obstante, a alternativa (C) apresentou coeficiente bisserial positivo (0,10), indicando que alguns estudantes de bom desempenho escolheram esta opção de resposta. Diante disso, seria prudente a revisão da alternativa (C).

Tabela 10 - Características psicométricas do Item 18

Estatisticas Clássicas															
ITEM 18															
ÍNDICES					PERCENTUAIS DE RESPOSTAS					COEFICIENTES BISSERIAIS					
ID	DISCR	ABAI	ACIM	BISE	A	B	C	D	E	A	B	C	D	E	
0.27	0.25	0.23	0.48	0.43	27,0	28,7	21,3	16,6	12,0	0,43	-0,18	-0,25	-0,11	-0,17	

Fonte: Produção própria, 2020

Gráfico 6. Item 18: Frequência de escolha do gabarito nos três grupos de desempenho (ABAI) (MEDI) e (ACIM)



Fonte: Produção própria (2020).

Neste item, o Descritor da matriz de referência da avaliação cobrado foi o D13 - *Resolver problema envolvendo a área total e/ou volume de um sólido (prisma, pirâmide, cilindro, cone, esfera)* - pertence ao eixo do conhecimento ou subescala "Grandezas e Medidas". Apresentou *Índice de Dificuldade* (ID) = 0,27, o que significa que apenas 27% dos estudantes acertaram esse item em particular, sendo, portanto, classificado como difícil.

A partir do gráfico 6, representativo do padrão de respostas do gabarito nos três grupos de desempenho, percebemos que houve um número bastante significativo (23%) de participantes do grupo de menor desempenho (ABAI) que acertaram o item. Não obstante, em relação ao grupo de maior desempenho (ACIM), onde se espera um percentual maior de acerto, esse número não foi expressivo: apenas 48%, conferindo ao item um *Índice de Discriminação D* igual a 0,25, indicando um item carente de revisão, pois não conseguiu distinguir bem os alunos com baixo e alto desempenho. No que diz respeito à *correlação ponto bisserial,* o gabarito (A) apresentou valor positivo (0,43), indicando um bom poder

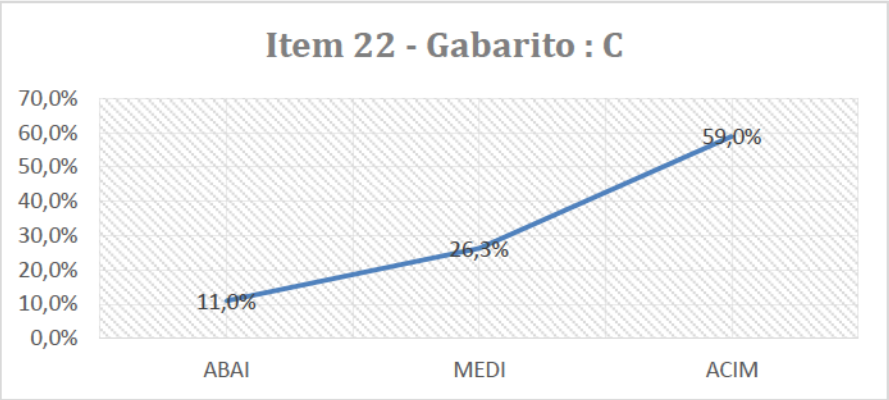
discriminativo. Em relação aos distratores, como esperado todos apresentaram valores negativos.

Tabela 11 - Características psicométricas do Item 22

Estatisticas Clássicas														
ITEM 22														
ÍNDICES					PERCENTUAIS DE RESPOSTAS					COEFICIENTES BISSERIAIS				
ID	DISCR	ABAI	ACIM	BISE	A	B	C	D	E	A	B	C	D	E
0.31	0.48	0.11	0.59	0.45	35,7	12,2	30,7	14,5	6,9	-0,19	-0,18	0,45	-0,11	-0,18

Fonte: Produção própria, 2020

Gráfico 7. Análise do Item 22: Frequência de escolha do gabarito nos três grupos de desempenho (ABAI) (MEDI) e (ACIM)



Fonte: Produção própria (2020).

No item 22, pertence à subescala "Grandezas e Medidas", o Descritor da matriz de referência da avaliação cobrado foi o D12 – Resolver problemas envolvendo o cálculo de figuras planas. Obteve-se ID = 0,30 sendo classificado como moderado ou mediano, seu *Índice de Discriminação D* foi igual a 0,48, indicando ser um bom item. Em relação aos coeficientes de correlação ponto bisseriais, o gabarito (C) apresentou valor positivo (0,45) como esperado, com elevado poder de discriminação, na mesma direção e sentido, os distratores apresentaram valores negativos, indicando que os alunos com bom desempenho não procuraram escolher essas alternativas.

4 APRESENTAÇÃO E DISCUSSÃO DOS RESULTADOS

Nesse tópico, apresenta-se um resumo dos principais resultados relativos às características psicométricas dos itens das subescalas de Matemática "Espaço e Forma e Grandezas e Medidas", tendo com foco de análise os itens da prova e não a prova como um todo, objetivando informar a adequação dos itens analisados à população testada. A confiabilidade do instrumento aplicado, também foi analisada, porém, nesse caso a base de análise foi a prova como um todo.

Para Erthal (2009), são dois os principais critérios utilizados para se verificar a adequação do item à população testada: quais sejam: o índice de dificuldade e o índice de discriminação ou poder discriminante. Segundo a autora, o índice de dificuldade propõe-se a

medir as diferenças individuais no que diz respeito ao rendimento alcançado no teste. Esse índice pode ser definido como sendo a razão entre a frequência de acertos no item e a quantidade de sujeitos que responderam a prova, assim, quanto menor o índice, maior a dificuldade.

4.1 Análise da dificuldade dos itens

O estudo realizado aponta – conforme exposto na Tabela 12, a seguir, que mostra os dados referentes ao Índice de Dificuldade dos itens analisados da primeira AGI, realizada em 2019 – que 57,2% dos itens foram classificados como moderados ou medianos e 42,8% podem ser classificados como difíceis.

Ainda em relação ao estudo sobre a dificuldade dos itens, observou-se que o grau de dificuldade médio dos itens da subescala "Grandezas e Medidas" possui um valor levemente superior ao de "Espaço e Forma", sugerindo itens mais difíceis. No entanto, em termos gerais, os itens da avaliação referentes às subescalas analisadas apresentaram grau de dificuldade mediano. Dessa maneira, é possível afirmar que, em média, o nível de dificuldade dos itens de Matemática relativos às subescalas analisadas foi considerado mediano ou moderado.

Tabela 12 – Taxa do *ID* dos 07 itens de Matemática, relativas às Subescalas Espaço e Forma e Grandezas e Medidas

Classificação	Taxa
Fácil	0,0%
Muito Fácil	0,0%
Moderado/Mediano	57,2%
Difícil	42,8%
Muito Difícil	0,0%
Total	100%

Fonte: Produção própria (2020).

4.2 Análise da discriminação dos itens

O Índice Discriminação - IDS, segundo Erthal (2009), estabelece relação entre escores totais altos ou baixos no teste e as respostas dadas a um item, sejam elas corretas ou incorretas. Desse modo o escore total no teste é usado para obter o Índice de Discriminação, assim sendo, conforme a autora, o critério para avaliar os itens é intrínseco ao próprio teste.

De acordo com a Tabela 13, a seguir, os dados nos indicam que 57,5% dos itens possuem boa capacidade de discriminação; são, portanto, bons itens. 28,5% são itens que necessitam de revisão e apenas um item, o que corresponde a 14% do total, relativos às subescalas de matemática em análise, foi considerado como ineficiente, e que poderia, portanto, ser excluído, tomando como base a Tabela de Ebel (1954). O item em questão é o de número 2, que apresentou simultaneamente alta dificuldade e baixa discriminação, no entanto, não chega a ser uma preocupação em relação aos resultados do teste.

Tabela 13 - Taxa do Índice de *Discriminação D* dos 07 itens de matemática relativas as Subescalas Espaço e Forma e Grandezas e Medidas

Índice de Discriminação D Classificação	Taxa
Satisfatório (item bom)	57,5%
Aceitável (sujeito a aprimoramento)	0,0
Item Marginal (necessita revisão)	28,5%
Ineficiente	14%
Total	100%

Fonte: Produção própria (2020).

Ainda com relação ao poder médio de discriminação dos itens, observa-se que os itens da subescala de "Espaço e Forma" mostraram um valor levemente superior ao de "Grandezas e Medidas", sugerindo itens mais discriminativos. Não obstante, em geral, os itens analisados se apresentaram com um bom poder discriminativo, portanto, com boa capacidade de diferenciar significativamente os indivíduos com desempenhos distintos na avaliação, garantindo uma boa consistência interna entre os itens e a totalidade da avaliação.

4.3 Confiabilidade da Avaliação Global Integrada - AGI

A TCT contribui, de modo singular, para a compreensão das propriedades psicométricas a partir de diversos coeficientes. Alguns, entre outros tantos, como os Índices de Dificuldade e Discriminação e o Coeficiente Alfa de Cronbach, foram objetos de análise ao longo desse trabalho. Nos próximos parágrafos nos deteremos à análise da confiabilidade da Avaliação Global Integrada – AGI através do Coeficiente Alfa de Cronbach.

Não é necessário grande empenho analítico para se concluir que avaliar a confiabilidade de um instrumento de medição utilizado em uma pesquisa é de extrema importância, na medida em que confere relevância para a mesma. O Alfa de Cronbach é uma técnica largamente utilizada para se mensurar a confiabilidade com base na consistência interna de um teste, verificando em boa medida, o quanto o conjunto de itens selecionados para o teste contribui para o escore total.

No que concerne à consistência interna dos itens analisados da Avaliação Global Integrada – AGI os dados nos revela que o instrumento aplicado apresentou alta confiabilidade. O valor de Alfa de Cronbach foi de 0,76, atendendo a referência sugerida por Freitas e Rodrigues (2005), que sugere que o valor mínimo aceitável em avaliações educacionais para o alfa é 0,70. Segundo os autores, valores do coeficiente entre 0,75 e 0,90 indicam alta confiabilidade.

CONCLUSÃO

Conforme demonstrado ao longo do trabalho, o presente estudo objetivou analisar os itens da Avaliação Global Integrada – AGI de matemática relativa às subescalas "Espaço e Forma" e "Grandezas e Medidas" pela Teoria Clássica dos Testes, centrando a análise nos itens da prova e não na prova como um todo, e assim contribuir para o melhor entendimento do desempenho dos alunos quando submetidos a esse tipo de prova.

Consideramos que a análise de itens nesse tipo de avaliação deve contemplar critérios estatísticos, que seja possível informar a adequação do item à população testada. Segundo Erthal (2009, p. 78) são dois os principais critérios com esse objetivo: o índice de dificuldade e o índice de discriminação ou poder discriminante.

Desse modo, o estudo em tela, avaliou propriedades psicométricas clássicas dos itens da Avaliação Global Integrada – AGI de matemática, relativas às subescalas "Espaço e Forma" e "Grandezas e Medidas", isto fica evidente, pela descrição numérica dos dados referentes à Dificuldade e Discriminação (vide gráficos e tabelas). Concluiu-se, nesta perspectiva, que o instrumento de avaliação utilizado possui um bom índice de confiabilidade em termos gerais, 86% dos itens analisados apresentaram qualidade psicométrica adequada para o contexto da avalição.

Face ao exposto, constatou-se a tendência de existir maior quantidade de itens com dificuldade mediana, embora em uma das subescalas, mais precisamente, "Grandezas e Medidas", o grau de dificuldade de seus itens foi considerado difícil.

Constatou-se também que, nas duas subescalas analisadas apenas o item 2 da subescala "Grandezas e Medidas", pode ser apontado como pouco discriminativo segundo o critério do *índice D (0,18)* ou da *correlação ponto bisserial (0,23)* que indica ser um item Ineficiente, portanto, deveria ser rejeitado.

Cabe ressaltar que, de acordo com a discriminação dos itens, proposta por Ebel (1954), os itens identificados como pouco discriminativos foram aqueles cujas *correlações ponto bisseriais ou polisseriais* e *índice D* foram menores que 0,30. Não obstante, a maioria dos itens analisados demonstra estarem adequados, discriminando satisfatoriamente os examinandos com desempenhos distintos na prova.

De toda sorte, os resultados do presente estudo possibilitarão aos usuários desta avaliação, principalmente, coordenadores pedagógicos e professores a conhecer a qualidade dos itens das subescalas analisadas em relação ao nível de dificuldade e à sua capacidade discriminativa, além da confiabilidade do teste aplicado como um todo.

Neste sentido, e por considerar a base de dados mencionada na introdução deste estudo, pode-se afirmar que esta pesquisa traz subsídios baseados em dados reais para uma discussão focada na análise de itens via Índices de Dificuldade e Discriminação, interpretados no cenário educacional da rede estadual de ensino do Piauí, contemplando o conteúdo matemático, relativos à 3ª série do ensino médio.

Por fim, como demonstrado ao longo desse trabalho, conclui-se que a Avaliação Global Integrada possui boa confiabilidade, apresentando coeficiente alfa com valor de 0,76, quanto aos Índices de Dificuldade e Discriminação, no geral os itens analisados apresentaram dificuldade mediana e boa capacidade de discriminação.

REFERÊNCIAS

ALMEIDA, D.; SANTOS, M. A. R. dos; COSTA, A. F. Aplicação do Coeficiente Alfa de Cronbach nos Resultados de um Questionário para Avaliação de Desempenho da Saúde Pública. In: ENCONTRO NACIONAL DE ENGENHARIA DE PRODUÇÃO, 30., 2010, São Carlos. **Anais Eletrônicos**.... São Carlos, 2010. Disponível em: < http://www.abepro.org.br/biblioteca/enegep2010_TN_STO_131_840_16412.pdf> Acesso em: 10 out. 2020.

ANDRADE, D. F.; TAVARES, H.R.; VALLE da C. **Teoria de Resposta ao item:** Conceitos e Aplicações. [S.I.]: Associação Brasileira de Estatística, São Paulo, 2000. 154p.

BORGATTO, A. F.; ANDRADE, D. F. Análise clássica de testes com diferentes graus de dificuldade. **Estudos em Avaliação Educacional**, São Paulo, v. 23, n. 52, p. 146-156, 2012.

ERTHAL, T. C. **Manual de Psicometria**. 8ª. Ed.: Rio de Janeiro; Jorge Zahar Ed., 200, 149p.

ERTHAL, T. C. **Manual de Psicometria**. 6.ed. Rio de Janeiro: Jorge Zahar, 2001, 149p.

FREITAS, A. L. P., RODRIGUES, S. G. A. Avaliação da confiabilidade de questionário: uma análise utilizando o coeficiente alfa de Cronbach In: SIMPÓSIO DE ENGENHARIA DE PRODUÇÃO, 12, 2005, Bauru-SP. Anais... Bauru-SP: UNESP, 2005.

GASPAR, I. A.; SHIMOYA, A. Avaliação da confiabilidade de uma pesquisa utilizando o Coeficiente alfa de Cronbach. In: Simpósio de Engenharia de Produção, 1, 2017, Catalão-Go. **Anais [...].**Catalão-Go: UFG, 2017. p. 1-7.

HORA, H. R. M. da; MONTEIRO, G. T. R., ARICA, J. Confiabilidade em Questionários para Qualidade*: Um Estudo com o Coeficiente Alfa de Cronbach.* **Produto & Produção**, vol. 11, n. 2, p. 85 – 103, jun. 2010. Disponível em: https://www.researchgate.net/publication/233793375_Confiabilidade_em_Questionarios_para_Qualidade_Um_Estudo_com_o_Coeficiente_Alfa_de_Cronbach/link/02bfe51006a53d1de3000000/download. Acesso em: 10 out. de 2020.

KNÜPFER, R. E. N.; AMARAL, A. do; HENNING, E. **Análise Clássica de Testes: Uma proposta de análise de desempenho dos estudantes na primeira fase da OBMEP II**. In: Colóquio Luso-brasileiro de Educação, 2. Joinville -SC. Anais **[...].**: Joinville -SC: UDESC, 2016. p. 272-283. Disponível em: https://revistas.udesc.br/index.php/colbeduca/article/view/8428 Acesso em 17 ago. 2020.

LORD, F. M. **Application of Item Response Theory to Practical Testing Problems. first. Hilsdale**,New Jersey, EUA: Lawrence Erlbaum Associates, 1980. 274 p.

MATTHIENSEN, Alexandre. Uso do Coeficiente Alfa de Cronbach em Avaliações por Questionários. **Documentos 48**, EMBRAPA: Boa Vista, RR, 2011. 31 p.

NAGEL, E. **Criação e validação de questões em estatística descritiva pela TCT e TRI**. 2018. Dissertação (Mestrado) - Programa de Pós-Graduação Profissional em Ensino de Ciências, Matemática e Tecnologias, Universidade do Estado de Santa Catarina, Santa Catarina, 2018.

PASQUALI, L. (Org). **Teoria e métodos de medida em Ciências do Comportamento**. Brasília: Laboratório de pesquisa em avaliação e medida/Instituto de Psicologia/UnB/INEP,1996, 432p.

PASQUALI, L.; PRIMI, R. Fundamentos da teoria de resposta ao item - tri**. Avaliação Psicológica**, v. 2, p. 99-110, 2003.

PASQUALI, Luiz. **Psicometria**: teoria dos testes na psicologia e na educação. 3. ed. Petrópolis, Rio de Janeiro: Vozes, 2009.

PASQUALI, Luiz. Psicometria – **Teoria dos testes na Psicologia e na Educação**. 5ed. Petrópolis: Vozes, 2013.

PITON-GONÇALVES, J.; ALMEIDA, A. M. Analise da dificuldade e da discriminação de itens de Matemática do ENEM. **REMAT**, Bento Gonçalves, RS, Brasil, v. 4, n. 2, p. 38-53, dezembro de 2018. Disponível em: https://periodicos.ifrs.edu.br/index.php/REMAT. Acesso em: 10 ago. 2020.

SILVEIRA, Fernando Lang da. **Considerações sobre o índice de discriminação de itens em testes educacionais**. Educação & Seleção, v. 7, 1983. Disponível em:https://www.researchgate.net/publication/318542967_Consideracoes_sobre_o_indice_de_discriminacao_de_itens_em_testes_educacionais. Acesso em: 14 set. 2020.

SOUSA, Leandro Araújo de. **Análise comparativa do Exame Nacional do Ensino Médio (ENEM) via Teoria Clássica dos Testes e Teoria de Resposta ao Item**. 2019. Tese doutorado – Programa de Pós-Graduação em Educação, Universidade Federal do Ceará, Fortaleza, 2019.

VILARINHO, A. P. L. Uma proposta de análise de desempenho dos estudantes e de valorização da primeira fase da OBMEP. 2015. Dissertação (Mestrado Profissional em Matemática) – Mestrado Profissional de Matemática, Universidade de Brasília, Brasília, 2015.