UNIVERSIDADE FEDERAL DO CEARÁ
FACULDADE DE EDUCAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM EDUCAÇÃO
DISSERTAÇÃO DE MESTRADO

LEANDRO ARAUJO DE SOUSA

ANÁLISE PSICOMÉTRICA DOS ITENS DE EDUCAÇÃO FÍSICA DO EXAME NACIONAL DO ENSINO MÉDIO (ENEM) VIA TEORIA CLÁSSICA DOS TESTES

FORTALEZA
2017

LEANDRO ARAUJO DE SOUSA

ANÁLISE PSICOMÉTRICA DOS ITENS DE EDUCAÇÃO FÍSICA DO EXAME NACIONAL DO ENSINO MÉDIO (ENEM) VIA TEORIA CLÁSSICA DOS TESTES

Dissertação apresentada ao Programa de Pós-Graduação em Educação da Universidade Federal do Ceará como requisito parcial para obtenção do título de Mestre em Educação. Área de concentração: Avaliação Educacional.

Orientadora: Prof.ª Dra. Adriana Eufrásio Braga.
Coorientador: Prof. Dr. Nicolino Trompieri Filho.

FORTALEZA

2017

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

S697a Sousa, Leandro Araujo de.
Análise psicométrica dos itens de educação física do Exame Nacional do Ensino Médio (ENEM) via teoria clássica dos testes / Leandro Araujo de Sousa. – 2017.
69 f. : il.

Dissertação (mestrado) – Universidade Federal do Ceará, Faculdade de Educação, Programa de Pós-Graduação em Educação, Fortaleza, 2017.
Orientação: Profa. Dra. Adriana Eufrásio Braga.
Coorientação: Prof. Dr. Nicolino Trompieri Filho.

1. Avaliação em larga escala . 2. Ensino Médio. 3. Teoria Clássica dos Testes. I. Título.
CDD 370

LEANDRO ARAUJO DE SOUSA

ANÁLISE PSICOMÉTRICA DOS ITENS DE EDUCAÇÃO FÍSICA DO EXAME NACIONAL DO ENSINO MÉDIO (ENEM) VIA TEORIA CLÁSSICA DOS TESTES

Dissertação apresentada ao Programa de Pós-Graduação em Educação da Universidade Federal do Ceará como requisito parcial para obtenção do título de Mestre em Educação. Área de concentração: Avaliação Educacional.

Aprovada em: 31 / 01 / 2017 .

BANCA EXAMINADORA

_____

Profa. Dra. Adriana Eufrásio Braga (Orientadora)

Universidade Federal do Ceará (UFC)

_____

Prof. Dr. Nicolino Trompieri Filho (Coorientador)

Universidade Federal do Ceará (UFC)

_____

Prof. Dr. José Airton de Freitas Pontes Junior

Universidade Estadual do Ceará (UECE)

A Deus.

Aos meus pais, Edmar e Auzenir.

A minha esposa, Grasiany.

AGRADECIMENTOS

A Profa. Dra. Adriana Eufrásio Braga e Prof. Dr. Nicolino Trompieri Filho, pela excelente orientação.

Ao professor participante da banca examinadora Dr. José Airton de Freitas Pontes Junior, pelo tempo despendido na análise da dissertação, pelas valiosas colaborações e sugestões.

Aos colegas da turma de mestrado, pelas reflexões, críticas e sugestões recebidas.

A minha família, pelo apoio e motivação na realização desse trabalho.

A minha esposa Grasiany, pela compreensão das ausências e por ser fonte de inspiração na realização dos objetivos acadêmicos.

“É necessário sempre pensar na avaliação no contexto de um processo formativo: a avaliação para orientar os procedimentos docentes; a avaliação para sugerir novas estratégias eficientes de ensino que levem a uma aprendizagem que seja relevante para o aluno como pessoa humana; a avaliação como um fator de orientação de todo o processo docente, envolvendo não apenas conhecimentos, mas incluindo o despertar de novos interesses e a formação de valores; a avaliação como uma ponte que une professor e aluno visando a um processo interativo gerador de novas aprendizagens; a avaliação como fator capaz de gerar elementos que facilitem a superação dos problemas curriculares e que muitas vezes decorrem de conflitos entre a realidade da escola e o contexto sociocultural em que a mesma se situa.”

Heraldo M. Vianna

RESUMO

Nos últimos anos tem crescido a importância das avaliações em larga escala no contexto brasileiro, com destaque nesse cenário o Exame Nacional do Ensino Médio (ENEM). Com sua reformulação em 2009, competências e habilidades da área de Educação Física têm sido inseridas na matriz de referência desse exame. Nesse mesmo ano é alterado também o método de análise dos resultados, realizado a partir da Teoria Clássica dos Testes (TCT), passando a ser utilizada a Teoria de Resposta ao Item (TRI), sob justificativa de ser mais adequada por permitir a comparabilidade dos resultados. Com isso, esta pesquisa objetivou analisar os itens de Educação Física do ENEM dos anos de 2009 a 2014 a partir da TCT. Para tanto, utilizou-se os microdados do exame disponibilizados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). Foram analisados os seguintes parâmetros métricos: validade, fidedignidade, dificuldade e discriminação. Utilizou-se como recurso o software SPSS, versão 20.0. Os itens apresentaram bons valores de correlação e adequação da amostra de itens. Apresentaram escores de comunalidade e cargas fatoriais inadequados para composição da prova. A Análise Fatorial Exploratória apresentou baixa explicação da variância considerando apenas um fator, mesmo a análise gráfica (*scree plot*) indicando a unidimensionalidade do teste. Os valores de fidedignidade da prova foram bons, não havendo influência dos itens de Educação Física. A dificuldade e discriminação apresentaram valores aceitáveis em quase todos os anos. No entanto, em 2014 a prova não apresentou unidimensionalidade, considerando a variância explicada, bem como na análise gráfica. Neste ano, os itens apresentaram alta dificuldade e baixa discriminação. Dessa forma, conclui-se que as provas de Linguagens e Códigos do ENEM apresentaram dificuldades de comprovação da unidimensionalidade, embora, tenha apresentado boa precisão, com exceção de 2014 e alguns itens de Educação Física do exame não apresentaram parâmetros adequados. Tais fatores podem comprometer a validade da medida e consequentemente dos resultados desse exame.

Palavras-chave: Avaliação em larga escala. Ensino Médio. Teoria Clássica dos Testes.

ABSTRACT

In recent years the importance of large-scale evaluations in the Brazilian context has grown, with emphasis in this scenario on the National High School Examination (ENEM). With its reformulation in 2009, the skills and abilities of the Physical Education have been inserted in the reference matrix of this exam. In that same year, the method of analysis of the results, based on the Classical Tests Theory (CTT), was also changed, using the Item Response Theory (IRT), under justification of being more adequate to allow the comparability of the Results. With this, this research aimed to analyze the Physical Education items of ENEM from the years 2009 to 2014 from the CTT. For that, we used the microdata of the exam provided by the National Institute of Studies and Educational Research Anísio Teixeira (INEP). The following metric parameters were analyzed: validity, reliability, difficulty and discrimination. SPSS software version 20.0 was used as a resource. The items presented good correlation values and adequacy of the item sample. They presented scores of commonality and factorial loads inadequate for the composition of the test. The Exploratory Factor Analysis presented low explanation of the variance considering only one factor, even the scree plot indicating that the test is unidimensionality. The reliability values of the test were good, with no influence of physical education items. The difficulty and discrimination presented values acceptable in almost every year. However, in 2014 the test did not present unidimensionality, considering the explained variance, as well as in the graphic analysis. This year, the items presented high difficulty and low discrimination. Thus, it is concluded that the Language and Codes tests of the ENEM presented difficulties in proving the unidimensionality, although it presented good accuracy, with the exception of 2014 and some Physical Education items of the exam did not present adequate parameters. Such factors may compromise the validity of the measure and consequently the results of such examination.

Keywords: Large-scale assessment. High school. Classical Theory of Tests.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

SAEB	Sistema de Avaliação da Educação Básica
ENEM	Exame Nacional do Ensino Médio
ENCCEJA	Exame Nacional e Certificação de Jovens e Adultos
SPAECE	Sistema Permanente de Avaliação da Educação Básica do Ceará
SIMAVE	Sistema Mineiro de Avaliação da Educação Pública
SINAES	Sistema Nacional de Avaliação da Educação Superior
PCN’s	Parâmetros Curriculares Nacionais
IES	Instituições de Ensino Superior
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
MEC	Ministério da Educação
PROUNI	Programa Universidade para Todos
SISU	Sistema de Seleção Unificado
FIES	Fundo de Financiamento Estudantil
TRI	Teoria da Resposta ao Item
TCT	Teoria Clássica dos Testes
LDB	Lei de Diretrizes e Bases da Educação
AFE	Análise Fatorial Exploratória
CCI	Curva Característica do Item
DIF	Funcionamento Diferencial do Item

LISTA DE SÍMBOLOS

%	Porcentagem
n	Número
r	Índice de correlação de pearson
gl	Graus de liberdade
X^2	Qui-quadrado
α	Coeficiente de precisão Alpha de Cronbach
d	Índice de dificuldade
r_{bp}	Coeficiente de correlação bisserial por ponto
***P* (r_{bp})**	Significância do coeficiente de correlação bisserial por ponto

ÍNDICE

1 INTRODUÇÃO

Avaliar é estabelecer comparações, contrastar uma situação real observada com uma referência, um ideal ou um paradigma para identificar diferenças que os afastam (SOUSA, 1994). As primeiras concepções de avaliação da aprendizagem estavam ligadas à ideia de medir, direcionada à avaliação do desempenho dos alunos (DEPRESBITERIS, 1989). Historicamente no Brasil, a avaliação tem sido realizada através de testes, entendida como provas objetivas, que a partir da década de 1960, quando são incorporadas às avaliações dos vestibulares, passam a ter significado e exercer alguma influência na concepção de avaliação em larga escala (VIANNA, 1985).

Atualmente, as avaliações não se limitam à verificação do desempenho dos estudantes, sendo também integradas aos programas como forma de melhorar a qualidade, ocorrendo com a nova orientação das avaliações a partir dos grandes investimentos na área de educação realizados por instituições financeiras, passando a exigir o diagnóstico dos efeitos dos empreendimentos (VIANNA, 1997).

A partir disso, a concepção de avaliação em larga escala aparece com foco direcionado para todo o sistema educacional, em que o objetivo é verificar a qualidade da educação (LIMA; GOMES; ANDRIOLA, 2014), possibilitando o direcionamento de políticas públicas para a sua melhoria. Assim, são estruturados sistemas nacionais de avaliação voltados para o Ensino Fundamental, Médio e Educação Superior.

Entre eles podemos encontrar: o Sistema de Avaliação da Educação Básica (SAEB), que busca diagnosticar o rendimento dos alunos do Ensino Fundamental e Médio; o Exame Nacional do Ensino Médio (ENEM) que avalia o desempenho de egressos e alunos dessa etapa e; o Exame Nacional e Certificação de Jovens e Adultos (ENCCEJA) que avalia esse público em nível de Ensino Fundamental (BRASIL, 2002a; 2002b; 2008). Além destes, os estados também elaboram seus sistemas de avaliação, destacando-se pelo pioneirismo: o programa de avaliação do desempenho da Rede Pública de ensino do Estado de Pernambuco em 1991; o Sistema Permanente de Avaliação da Educação Básica do Ceará (SPAECE) em 1992 e; o Sistema Mineiro de Avaliação da Educação Pública (SIMAVE), surgido em 1992 (HORTA NETO, 2007). Encontra-se também o Sistema Nacional de Avaliação da Educação Superior (SINAES) dedicado a esse nível de ensino e instituído em 2004, tendo o objetivo de

assegurar processo nacional de avaliação das instituições de educação superior dos cursos de graduação e do desempenho acadêmico de seus estudantes (BRASIL, 2004).

As avaliações nacionais em larga escala da Educação Básica concentram esforços nas disciplinas de português e matemática. A exceção é o ENEM e ENCCEJA que avaliam todas as disciplinas do Ensino Médio, embora contemple mais itens das disciplinas de português e matemática. A Educação Física foi inserida na matriz de referência do ENEM em 2009. Desde então foi possível identificar pelo menos 18 questões relacionadas ao objeto de estudo da Educação Física (FERNANDES; RODRIGUES; NARDON, 2013). Com isso, foi fortalecida a necessidade de avaliar cognitivamente os alunos desse nível de ensino nessa disciplina.

1.1 Justificativa

As problemáticas que permeiam a avalição educacional, como a seleção de instrumentos e o uso dos resultados, têm sido amplamente discutidas pelos pesquisadores, estudiosos, professores e pessoas ligadas direta ou indiretamente à avaliação. Em Educação Física esse tema é ainda mais polêmico por causa da sua multiplicidade de objetivos ligados aos aspectos conceituais, procedimentais e atitudinais, exigindo uma variedade de instrumentos que contemplem todos esses fatores.

Quando se trata de avaliação cognitiva em Educação Física, percebe-se certa resistência de alguns professores e pesquisadores da área. Todavia, a partir da elaboração dos Parâmetros Curriculares Nacionais (PCN's), a dimensão cognitiva do ensino de Educação Física ganhou destaque, uma vez que, o documento ressalta o desenvolvimento de competências e habilidades cognitivas específicas para esse componente curricular (BRASIL, 2000). Essas medidas podem ter contribuído para que os conteúdos da Educação Física fossem incluídos na matriz de referência do ENEM a partir de 2009, desde então, itens relacionados a essa disciplina têm sido contemplados nessa avaliação.

Por ser recente, estudos que se debruçam na análise da Educação Física em avaliações de larga escala ainda são escassos. Encontram-se estudos que analisam as influências da inserção da Educação Física na matriz de referência do ENEM nos currículos e na prática dos professores da Educação Básica (BELTRÃO, 2014); que discutem a "desvantagem" dos estudantes do turno noturno, já que alguns são isentos da disciplina por determinação legal (FERNANDES; RODRIGUES; NARDON, 2013) e que analisam

qualitativamente os itens de Educação Física desse exame (SOUZA JÚNIOR et al., 2012). No entanto, estudos que analisam especificamente os aspectos psicométricos dos itens de Educação Física do ENEM são insuficientes ou mesmo inexistentes.

A partir disso, o estudo possibilitará conhecer as características psicométricas dos itens de Educação Física do ENEM a partir das respostas dos candidatos ao exame, permitindo assim, conhecer seus parâmetros métricos, adequabilidade e qualidade desses itens em medir conhecimentos e consequentemente estimar o desempenho nessa área dos candidatos que se submetem ao exame. Tal análise se justifica desde que, nos últimos anos, o ENEM passou a ser utilizado como forma de seleção dos candidatos aos cursos superiores das Instituições de Ensino Superior (IES) públicas e privadas brasileiras. Com isso, torna-se necessário a utilização de um instrumento confiável para realização justa dessas seleções.

1.2 Problema de pesquisa

No Brasil, as avaliações em larga escala passaram a ser discutidas na década de 1980 (GATTI, 2013). Isso fez com que fossem desenvolvidos estudos sobre sistemas educacionais, escolas, rendimentos dos alunos, insumos e oportunidades educacionais (FREITAS, 2013). A partir disso, apesar do desenvolvimento ocorrido nessa área, ainda há muito a se investigar.

Quando se considera estudos sobre a avaliação em larga escala em Educação Física no Brasil, poucas pesquisas são encontradas (SOUZA JÚNIOR et al., 2012; FERNANDES; RODRIGUES; NARDON, 2013; BELTRÃO, 2014). Entretanto, esses estudos apresentam caráter amplo do assunto e sem aprofundamentos nas análises dos dados das avaliações. A partir disso, propomos uma análise psicométrica mais detalhada dos itens de Educação Física do Exame Nacional do Ensino Médio.

As características psicométricas dos itens do ENEM são pouco exploradas nos estudos e pesquisas em avaliação educacional, e quando se trata dos itens de Educação Física as pesquisas são quase inexistentes, salvo os relatórios pedagógicos que traçam uma breve e superficial análise, avaliando o comportamento dos participantes em relação a cada item (INEP, 2013). Diante disso, surge a seguinte pergunta: Os itens de Educação Física do ENEM apresentam características psicométricas adequadas para medir os conhecimentos dos candidatos nessa área?

1.3 Objetivos

1.3.1 Geral

Analisar características psicométricas dos itens de Educação Física do Exame Nacional do Ensino Médio (ENEM) de 2009 a 2014 via Teoria Clássica dos Testes.

1.3.1 Específicos

- ✓ Estimar os parâmetros de dificuldade e discriminação dos itens de Educação Física do ENEM;

- ✓ Verificar a adequabilidade dos itens de Educação Física para a validade e fidedignidade da prova de Linguagens e Códigos do exame.

2 EDUCAÇÃO FÍSICA EM AVALIAÇÕES DE LARGA ESCALA: O CASO DO EXAME NACIONAL DO ENSINO MÉDIO – ENEM[1]

2.1 Origem e concepção das avaliações em larga escala no Brasil

As primeiras medições em educação no Brasil remontam ao início do século XX, em que buscavam levantar dados sobre o quantitativo de escolas, professores, matrículas, repetências (HORTA NETO, 2007). Entretanto, já no século XIX, nos Estados Unidos, ocorreram experiências na elaboração de um sistema de testagem para avaliar em larga escala com finalidade de sugerir melhorias para a educação (DESPREBITERIS, 1989).

As discussões sobre avaliações em larga escala e a elaboração de um projeto piloto no Brasil, visando à qualidade da educação, ocorreram a partir da década de 1980 (WAISELFISZ, 1991; GATTI, 2013). No entanto, foi na década de 1990 que as avaliações passam a ser utilizadas em diferentes níveis administrativos, buscando soluções para alguns problemas educacionais e na tentativa de que seus resultados pudessem determinar o bom desempenho dos estudantes (VIANNA, 2003). Todavia, para o autor, as avaliações tão somente fazem o diagnóstico dos problemas, mas por si só, não os solucionam, sendo necessário traçar outros caminhos.

As problemáticas das avaliações na Educação Básica ficaram mais evidentes por conta da ampla divulgação dos resultados centralizados no rendimento dos alunos e no desempenho dos sistemas de ensino que foram insatisfatórios (COELHO, 2008). Atualmente, as avaliações de larga escala são consideradas um procedimento necessário para formulação e monitoramento de políticas educacionais, tendo em vista uma gestão para aprimoramento de sua qualidade (LIMA, 2012). A partir disso, sugiram sistemas de avaliação de abrangência nacional, estadual e municipal e nos diferentes níveis de ensino com finalidade de atender essas demandas. Neste texto será dada ênfase ao ENEM.

2.2 O Exame Nacional do Ensino Médio – ENEM

A partir da expansão dos sistemas de avaliação em larga escala nos diversos níveis de ensino, o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira

[1] Texto publicado na Revista Educação & Linguagem, v. 2, n. 1, jun., 2015.

(INEP) cria o ENEM que é instituído pela Portaria do Ministério da Educação (MEC) Nº 438 de 28 de maio de 1998 para ser aplicado anualmente aos egressos e estudantes no final do Ensino Médio, com a finalidade de avaliar os conhecimentos necessários para o exercício da cidadania (BRASIL, 1998). De acordo com a referida lei, os objetivos do exame estão voltados para a autoavaliação e como processo alternativo para o ingresso no mercado de trabalho, ensino profissionalizante e Educação Superior.

Com essas características, o processo teve pouca participação nas suas primeiras edições, chegando a pouco mais que 150 mil inscritos na primeira edição. No entanto, com algumas reformulações que ocorreram, entre elas: a criação do Programa Universidade para Todos (PROUNI) em 2004, na qual oferece bolsa para estudantes a partir das notas do ENEM (BRASIL, 2004); a implementação do Sistema de Seleção Unificado (SISU) em 2010 com a utilização das notas do ENEM a partir de 2009 (BRASIL, 2010a), em que grande número de instituições passa a adota-lo como forma de ingresso em seus cursos (VIGGIANO; MATTOS, 2013); a obrigatoriedade da nota do ENEM para a contratação do Fundo de Financiamento Estudantil (FIES) em 2010 (BRASIL, 2010b), o número de inscritos no exame teve grande aumento, chegando a mais de 8,7 milhões de inscritos em 2014, segundo dados do INEP.

O ENEM é uma prova constituída de itens de múltipla escolha e uma redação, na qual busca avaliar competências e habilidades desenvolvidas pelos alunos no decorrer da Educação Básica, sendo orientada por uma matriz de referência construída especificamente para o exame (BRASIL, 2002a).

A matriz de referência atualmente é dividida em quatro grandes áreas do conhecimento: Linguagens, Códigos e suas Tecnologias, que contempla os conhecimentos de Português, Educação Física, Artes, Língua Estrangeira Moderna e Tecnologia da Informação e Comunicação; Ciências da Natureza e suas Tecnologias, que abrange a Biologia, Química e Física; Ciências Humanas e suas Tecnologias, envolvendo a História, Geografia, Sociologia e Filosofia e; Matemática e suas Tecnologias (INEP, 2013).

Para aprimorar a técnica de análise dos resultados, a partir de 2009 a Teoria da Resposta ao Item (TRI) foi incorporada ao exame. Para Corti (2013) e Gatti (2013) com essa inovação foi possível tornar os resultados comparáveis, tanto das diferentes populações, contextos ou com resultados anteriores, embora com alguns questionamentos. Segundo Viggiano e Mattos (2013) com a TRI é possível considerar que duas avaliações distintas que

sejam construídas com base nas mesmas competências e habilidades de uma matriz de referência sejam equivalentes.

2.3 Avaliação em Educação Física

Historicamente, o ensino de Educação Física esteve vinculado às tendências militaristas, esportivistas e higienistas, em que as práticas avaliativas se pautavam nos testes de aptidão física, prevalecendo a análise dos níveis de força, velocidade, resistência cardiorrespiratória e outras capacidades físicas dos alunos (PONTES JUNIOR; TROMPIERI FILHO, 2011).

A partir da década de 1980, com a redemocratização do Estado brasileiro, surgem novas abordagens que orientam a Educação Física em contraponto às concepções ora vigentes (DARIDO, 2011), influenciando também nas práticas avaliativas. Na segunda metade da década de 1990, com a promulgação da Lei de Diretrizes e Bases da Educação – LDB (BRASIL, 2013) e com a inserção da Educação Física como componente curricular das escolas, são elaborados documentos de orientação para a disciplina, os Parâmetros Curriculares Nacionais – PCN's (BRASIL, 2000).

Os PCN's de Educação Física, além das dimensões motora e sócio-afetiva, enfatizam competências e habilidades cognitivas. Com isso, os aspectos teóricos antes esquecidos passam a ganhar importância, em que instrumentos como provas são agregados no processo de avaliação como mostra os estudos de Pontes Junior, Soares e Trompieri Filho (2014a), que ao analisar a perspectiva discente sobre os instrumentos de avaliação em Educação Física, constatou-se que trabalhos e provas escritas estão entre os instrumentos mais utilizados nas aulas dessa disciplina.

Ao avaliar é necessário definir os objetivos. Em Educação Física os objetivos educacionais estão ligados às capacidades físico-esportivas, cognitivas e sócio-afetivas (PONTES JUNIOR; SOARES; TROMPIERI FILHO, 2014b). Precisa-se observar também que os objetivos educacionais devem contemplar toda diversidade de comportamentos a serem desenvolvidos durante o programa educacional. Para tanto, Benjamin S. Bloom e Associados propõem uma Taxonomia dos Objetivos Educacionais dividida nos domínios cognitivo, afetivo e psicomotor (BLOOM et al., 1956).

Para Arslan, Erturan e Demirhan (2013) entre os objetivos da Educação Física está o desenvolvimento da aptidão física e do estilo de vida ativo e saudável, como também

aspectos afetivos, cognitivos, sociais e psicomotores dos alunos, em que esses devem ser medidos e avaliados. Segundo os autores, entre os desafios dos professores de Educação Física está a desenvolvimento de avaliações significativas que analisem o progresso dos alunos na realização dos objetivos educacionais.

2.4 Educação Física em avaliações de larga escala

As avaliações em larga escala ganharam destaque nas discussões educacionais no Brasil a partir da década de 1980 (GATTI, 2013), fazendo com que fossem desenvolvidos estudos nessa área, mas ainda, apesar do desenvolvimento ocorrido nessa área, ainda há muito a se investigar, uma vez que as problemáticas tais como a seleção de instrumentos e o uso dos resultados, têm sido amplamente discutidas e com pouco consenso entre os pesquisadores da área. Em Educação Física esse tema é ainda mais polêmico por causa da sua multiplicidade de objetivos.

Poucos estudos foram empreendidos na análise da Educação Física em avaliações de larga escala. As poucas pesquisas na área se limitam a uma discussão teórica do assunto, vinculando-as a questões relacionadas aos impactos das avaliações na Educação Básica (BELTRÃO, 2014; FERNANDES; RODRIGUES; NARDON, 2013).

Entre os estudos que envolvessem a temática no Brasil encontra-se o de Pontes Junior, Trompieri Filho e Almeida (2014), em que a partir da perspectiva de professores e pesquisadores de Educação Física no Ensino Fundamental foi desenvolvida e validada uma matriz de referência para avaliação cognitiva em larga escala dos conteúdos da Educação Física (9º ano do Ensino Fundamental). A matriz contempla três dimensões: dimensão sociocultural das práticas corporais; dimensão biológico-funcional da atividade física e; dimensão técnico-competitiva dos esportes, podendo ser usada como parâmetro para avaliação do referido nível de ensino.

Com a consolidação da Educação Física como componente curricular obrigatório da Educação Básica (BRASIL, 2013) e os aspectos cognitivos da disciplina ressaltados nos PCN’s (BRASIL, 2000), passaram a ter uma preocupação com os conhecimentos adquiridos pelos alunos a cerca da área.

Com isso, a partir da elaboração dos Parâmetros Curriculares Nacionais (PCN’s), a dimensão cognitiva do ensino de Educação Física ganhou destaque, uma vez que, o documento ressalta o desenvolvimento de competências e habilidades cognitivas específicas

para esse componente curricular (BRASIL, 2000), conteúdos da Educação Física foram incluídos na Matriz de Referência do ENEM em 2009.

A partir disso, os conhecimentos da disciplina passam a ser requisitados na resolução da prova e desde então já foram aplicadas seis provas em que duas foram canceladas por vazamento de conteúdo, sendo que nestas, foi possível identificar pelo menos 18 itens que tinham conteúdos relacionados com a Educação Física (FERNANDES; RODRIGUES; NARDON, 2013).

A Educação Física está inserida na Matriz de Referência do ENEM na área de Linguagens, Códigos e suas Tecnologias. Isso ocorre porque a disciplina trabalha com linguagem corporal. Segundo Santos, Marcon e Trentin (2012) isso se explica porque o uso da linguagem do corpo possibilita e estimula a comunicação das diferentes culturas, em que os alunos interagem com a cultura corporal.

> **Competência de área 3** - Compreender e usar a linguagem corporal como relevante para a própria vida, integradora social e formadora da identidade.
>
> **H9** - Reconhecer as manifestações corporais de movimento como originárias de necessidades cotidianas de um grupo social.
>
> **H10** - Reconhecer a necessidade de transformação de hábitos corporais em função das necessidades cinestésicas.
>
> **H11** - Reconhecer a linguagem corporal como meio de interação social, considerando os limites de desempenho e as alternativas de adaptação para diferentes indivíduos (BRASIL, 2015).

As referidas habilidades estão relacionadas a um conjunto de conteúdos que são necessários para desenvolvê-las e são denominadas de objetos de conhecimento associados à Matriz de Referência.

> **Estudo das práticas corporais: a linguagem corporal como integradora social e formadora de identidade** – performance corporal e identidades juvenis; possibilidades de vivência crítica e emancipada do lazer; mitos e verdades sobre os corpos masculino e feminino na sociedade atual; exercício físico e saúde; o corpo e a expressão artística e cultural; o corpo no mundo dos símbolos e como produção da cultura; práticas corporais e autonomia; condicionamentos e esforços físicos; o esporte; a dança; as lutas; os jogos; as brincadeiras (BRASIL, 2015).

Beltrão (2014) reconhece a importância do ENEM como um instrumento de diagnóstico da educação, possibilitando a definição de políticas públicas, todavia, admite que a inserção da Educação Física nesse e em outros exames pode influenciar a prática docente,

em que as aulas podem se resumir a preparação para a prova. Assim, segundo o autor, a disciplina passaria a teorizar sobre as manifestações da cultura corporal e não mais praticá-las. Mas é importante ressaltar também que os aspectos teóricos referentes aos conteúdos também são importantes, uma vez que poderá levar os alunos a tomarem consciência do contexto em que estas manifestações estão inseridas, permitindo uma prática crítica e reflexiva.

A Educação Física enquanto parte de avaliações de larga escala é recente. No entanto, já exerce alguma influência na prática dos professores de Educação Básica, em que esses já incorporam aulas teóricas e realizam avaliações do conhecimento adquiridos pelos alunos nessa disciplina.

3 PSICOMETRIA EM AVALIAÇÃO EDUCACIONAL: TEORIA CLÁSSICA DOS TESTES E TEORIA DE RESPOSTA AO ITEM

Os testes e resultados obtidos nas avaliações em larga escala têm sido analisados através de técnicas psicométricas, pois como afirma Sartes e Souza-Formigoni (2013), instrumentos e testes construídos com base nessa técnica tem sido uma forma de avaliar objetivamente os fenômenos psicológicos. Segundo Pasquali (2009a) ”a psicometria procura explicar o sentido que têm as respostas dadas pelos sujeitos a uma série de tarefas, tipicamente chamadas de itens”. Apesar de se fundamentar em uma base epistemológica eminentemente quantitativista, assumindo pressupostos da teoria da mensuração, a psicometria é considerada como um ramo da psicologia, ou seja, das ciências empíricas (PASQUALI, 2009b).

Dessa forma, a psicometria faz uso de testes, em que pesquisadores das ciências psicossociais utilizam-se destes para estimar o comportamento psicológico do sujeito, prevendo de certa forma o erro presente nessa medida (MUÑIZ, 1998). Para tanto, indagando sobre as condições para se obter uma medida adequada, o autor ressalta a observância de três características: a fiabilidade, denominada como o grau de precisão na medição do teste; a validade, ou seja, a garantia que as inferências realizadas a partir da medição são corretas e; a fundamentação teórica em que o teste está embasado.

Entre as técnicas utilizadas para analisar os resultados temos a Teoria Clássica dos Testes (TCT), apresentados através de scores padronizados, utilizando-os como processo de seleção das pessoas (VALLE, 2000). A TCT se preocupa com o resultado final através do somatório dos escores dos itens de um teste, expresso no escore total, tendo interesse em produzir testes válidos e com qualidade (PASQUALI, 2009a). No entanto, a TCT não permite analisar os dados entre amostras diferentes, assim é muito difícil comparar pessoas que não foram submetidas à mesma avaliação. Isso ocorre porque os resultados são dependentes do grupo avaliado e dos itens, sendo que a análise é realizada em função do instrumento como um todo (ANDRADE; TAVARES; VALLE, 2000).

Para resolver esses problemas da TCT, foi desenvolvida a Teoria da Resposta ao Item (TRI), que “é um conjunto de modelos matemáticos que considera o item como unidade básica de análise” (ANDRADE; LAROS; GOUVEIA, 2010). Dessa forma, a TRI não está interessada no escore total de um teste, mas sim em cada item de um teste, tendo interesse em produzir itens válidos e com qualidade (PASQUALI, 2009a). A vantagem da TRI, segundo

Andrade, Tavares e Valle (2000), é que ela permite comparar populações diferente, desde que o instrumento tenha alguns itens em comum, e mesmo indivíduos da mesma população ainda que submetidos a provas distintas. Segundo os mesmos autores alguns problemas em educação podem ser solucionados, como por exemplo, permitir o acompanhamento de uma série ao longo dos anos, bem como a comparação de desempenho entre escolas.

A seguir será detalhada cada teoria, TCT e TRI, apresentando os parâmetros estatísticos de qualidade dos itens, tais como: unidimensionalidade, validade, fidedignidade, dificuldade, discriminação, acerto casual e vieses dos itens para a primeira; unidimensionalidade, independência local, dificuldade, discriminação, acerto casual e vieses para a segunda.

3.1 Teoria Clássica dos Testes (TCT)

A TCT tem interesse e trabalha unicamente enquanto critério (aptidão ou aquilo que o teste deve medir) os comportamentos (variáveis observáveis), podendo ser entendido como o escore de um teste, que podem ser presentes ou futuros, ou seja, realidades físicas, por isso se diz que ela trabalha ao nível do tau (τ) do ser humano (PASQUALI, 2009b).

Esse modelo foi inicialmente desenvolvido por Spearman (MUÑIZ, 1998; PASQUALI, 2009b) e axiomatizada por Gulliksen, segundo o qual distingue três elementos: o escore empírico (T), o escore verdadeiro (V) e o erro (E), em que segundo Pasquali (2009b) se define como:

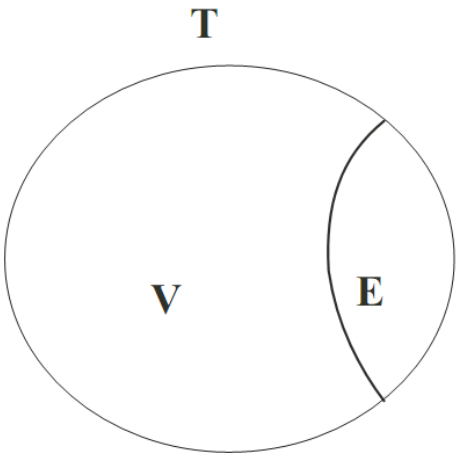
T = escore bruto ou empírico do sujeito, que é a soma dos pontos obtidos no teste;

V = escore verdadeiro, que seria a magnitude real daquilo que o teste quer medir no sujeito e que seria o próprio T se não houvesse erro de medida;

E = o erro cometido nesta medida.

Tendo isso, a TCT dispõe de alguns pressupostos básicos que são pilares dessa teoria. Aqui nos deteremos ao postulado fundamental: T = V + E, ou seja, o escore empírico é a soma do escore verdadeiro mais o erro (PASQUALI, 2009b). Na Figura 1 esse postulado é apresentado de forma esquemática.

Figura 1 – Componentes do escore T.



Fonte: Pasquali, 2009b.

Da mesma forma, o erro é o escore empírico subtraído do escore verdadeiro (E = T – V), assim como, o escore verdadeiro é o escore empírico subtraído do erro (V = T – E). Segundo Pasquali (2009b), o erro de medida está presente em qualquer operação empírica. Essa afirmação pode ser fundamentada em Popper (1972), quando trata da precisão nas medidas. O autor coloca que a medição consiste na coincidência entre dois pontos, o do instrumento e o do objeto medido. No entanto, ressalta ainda que essa coincidência não ocorre efetivamente, argumentando que não há uma "fusão" entre os dois pontos, mas sim uma justaposição, uma proximidade, ou seja, que o objeto se colocou entre os pontos da escala de um instrumento de forma ótima, apresentando níveis aceitáveis de erro.

Assim, segundo Pasquali (2009b), o objetivo da TCT é dispor de técnicas estatísticas que visem controlar ou predizer o tamanho do erro. Esses podem ocorrer por influência de muitos fatores, como apresenta Campbell e Stanley (1979), entre outros:

a) Erros de testagem: ocasionadas por má aplicação de um teste;
b) Erros do instrumento: tanto por mudanças na sua calibração como por mudanças nos observadores que podem atribuir escores diferentes a uma mesma situação, alterando as medidas produzidas;
c) Vieses de amostragem: devido a seleção diferencial, ou seja, formação de grupos não homogêneos;
d) Vieses do sujeito (maturação): condições internas dos respondentes, tais como indisposição física e psicológica;

e) Fatores ambientais (efeitos reativos de condições experimentais): exposição a condições que interfiram nas ações dos respondentes.

3.1.1 Parâmetros dos itens: dificuldade, discriminação e acerto casual

3.1.1.1 Índice de dificuldade

A dificuldade dos itens segundo a TCT é definida como a proporção de sujeitos que respondem corretamente ao item. Dessa forma, quanto mais sujeitos erram determinado item, mais difícil ele é. Por vezes é utilizada a proporção de sujeitos que acertam o item, sendo denominado, dessa forma, de índice de facilidade (VIANNA, 1976). Com isso, a dificuldade do item só é utilizada em contexto de testes de aptidão, em que há apenas respostas certas e erradas (PASQUALI, 2009b). Por julgar mais adequada a posição de Vianna (1976), adaptou-se a fórmula apresentada por Pasquali (2009b) a partir de sua concepção:

$$ID = \frac{E}{N}$$

em que,

ID = índice de dificuldade;

E = número de sujeitos que erraram o item;

N = numero total de sujeitos que responderam o item.

Suponha-se, dessa forma, que um determinado teste foi respondido por um grupo de 40 alunos. O item 1 foi respondido incorretamente por 10 alunos e o item 2 por 30 alunos. Aplicando, ficaria da seguinte forma:

$$ID\ (i\ 1) = \frac{10}{40} = 0{,}25 \qquad\qquad ID\ (i\ 2) = \frac{30}{40} = 0{,}75$$

Portanto, diz-se o item 1 ser mais fácil do que o item 2. Todavia, é necessário saber que a dificuldade do item é sempre dependente do conjunto de sujeitos examinados, e que apesar de se tentar estimar sua dificuldade, essa pode não corresponder à dificuldade real

do item, podendo ou não ser confirmada após a sua aplicação (VIANNA, 1976). Recomenda-se ainda que em testes de escolaridade disponha-se de itens que estejam em uma faixa de dificuldade entre 0,20 a 0,80 com índice médio entre 0,40 e 0,60, fora disso, os itens podem gerar uma série de problemas (VIANNA, 1976).

3.1.1.2 Índice de discriminação

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 (PASQUALI, 2009b; VIANNA, 1976). O índice de discriminação informa a coerência dos escores do item com os escores do teste, assim, quanto maior for o índice dos itens, maior será a homogeneidade do teste (SILVEIRA, 1983). O cálculo da discriminação do item pode ser obtido de duas formas (PASQUALI, 2009b): 1) pelos grupos-critério e 2) pela correlação do item com o escore total dos itens (correlação item-total).

Em casos de testes de desempenho escolar utiliza-se como critério a diferença entre os 27% do grupo que obtiveram os melhores resultados e os 27% dos sujeitos com os resultados mais baixos. Recomenda-se 27% quando utilizado para uma amostra grande e 30% para amostras pequenas (PASQUALI, 2009b).

Assim, se 80% do grupo superior e 50% do inferior acertaram determinado item, significa que esse item tem índice de discriminação 30. Também pode ser dado em proporção, nesse caso os valores são 0,80 e 0,50, respectivamente, com índice discriminativo de 0,30. Esse valor é denominado de índice D e varia de 0 a 100 se for calculado em porcentagem ou de 0 a 1 se em proporção (VIANNA, 1976).

No entanto, Silveira (1983) apresenta algumas restrições a esse procedimento: 1) é dependente da extensão (tamanho) dos grupos extremos, pois se a proporção dos grupos for menos ou maior que 27%, o índice de discriminação tenderá a aumentar ou diminuir, respectivamente; 2) depende do escore máximo no item, pois se dois itens do mesmo teste tiverem escores diferentes terão índices diferentes e; 3) depende do desvio padrão dos escores do item, pois mesmo que dois itens de um teste tenham escores idênticos, esses terão índices diferentes se os desvios padrões forem diferentes, o que tiver maior desvio padrão será mais discriminativo. Dessa forma, recomenda-se a correlação item-total como melhor medida de

consistência do escore do item com o escore do teste, ou seja, é a melhor medida de discriminação dos itens (SILVEIRA, 1983).

O índice de discriminação a partir da correlação item-total pode ser obtido de várias formas (correlação bisserial e bisserial por ponto, correlação tetracórica, correlação phi). No entanto, nos deteremos especialmente a uma: a correlação bisserial por ponto, por entender ser a mais adequada quando se trata de testes de rendimento escolar, objeto desse estudo.

A correlação bisserial por ponto é utilizada quando uma das variáveis (nesse caso o item) é dicotômica discreta, ou seja, quando só há duas possibilidades de resposta, não havendo continuidade subjacente entre essas categorias (FIELD, 2009), por exemplo, certo e errado, sim e não. Em testes de rendimento, mesmo utilizando itens de múltipla escolha, só há duas possibilidades, acertar ou errar a resposta, a não ser que exista mais de uma alternativa correta, o que dificilmente ocorre nesse tipo de teste. Dessa forma, ao analisar os itens, eles são dicotomizados, em que 0 corresponde a errado e 1 a certo. O modelo matemático dessa técnica é dado da seguinte forma (PASQUALI, 2009b):

$$r_{pb} = \frac{\overline{X}_A - \overline{X}_T}{S_T}\sqrt{\frac{p}{q}}$$

em que,

r_{pb} = correlação bisserial por ponto;

$\overline{X}_A$ = 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;

q = 1 – p.

No entanto, quando se propõem diferenciar sujeitos de características específicas, tais como masculino e feminino, classe econômica baixa e alta, alunos de escola diurna e noturna, se calcula a diferença das médias desses grupos. Assim, para um resultado mais

apurado da diferença real entre os dois grupos utiliza-se o Teste "t" de Studant para amostras independentes. Como esse teste utiliza-se das médias e variâncias, é necessário que os escores constituam uma variável contínua e que nenhum dos grupos possua variância zero (PASQUALI, 2009b). Adaptada para a presente situação, a fórmula desse teste é dada da seguinte forma (FIELD, 2009):

$$t = \frac{\bar{X}_{S-}\bar{X}_{S}}{\sqrt{\frac{S_S^2}{n_S} + \frac{S_I^2}{n_I}}}$$

em que,

t = teste "t" de Student;

$\bar{X}_{S\,e}\bar{X}_{S}$ = médias dos grupos superior e inferior;

$S_S^2\ e\ S_I^2$ = variâncias dos grupos superior e inferior;

$n_S e\ n_i$ = número de sujeitos nos grupos superior e inferior com graus de liberdade $(n_S - 1\ e\ n_i - 1)$.

3.1.1.3 Acerto casual

Ao realizar um teste de desempenho escolar pode ocorrer de o respondente acertar o item mesmo não sabendo da resposta. Essa é uma questão de probabilidade. Digamos que um sujeito responda aleatoriamente um item de múltipla-escolha com cinco alternativas. Espera-se que ele tenha uma probabilidade de 20% de acertar o item. Se o item tem quatro alternativas, a probabilidade de acerto será de 25%, pois cada item terá a mesma chance de ser escolhido, caso tenha se dado aleatoriamente. Dessa forma, o item poderá ter sua dificuldade afetada. A fórmula que considera essa situação é a seguinte (PASQUALI, 2009b; BONILLO, 2013):

$$ID = \frac{A - \frac{E}{K-1}}{N}$$

em que,

ID = índice de dificuldade;

A = número de sujeitos que acertam o item;

E = número de sujeitos que erram o item;

K = número de alternativas de resposta ao item;

N = número total de sujeitos.

Como exemplo, suponhamos que um dado item de múltipla-escolha com cinco alternativas foi respondido por 100 alunos, dos quais 70 acertaram o item. Se fosse realizado o índice de dificuldade (nesse caso índice de facilidade, pois se está considerando os sujeitos que acertaram o item) sem considerar o acerto casual, este item teria dificuldade em proporção de 0,70 (70/100. Ver fórmula do índice de dificuldade). No entanto, considerando a possibilidade de acerto casual, o item torna-se mais difícil (ID = 0,63).

$$ID = \frac{70 - \frac{30}{5-1}}{100} = \frac{62,5}{100} = 0,63$$

Esse parâmetro pressupõe, por tanto, que poderá haver respostas corretas conseguidas por acaso para cada K – 1 respostas incorretas. Essa suposição é alvo de muitas críticas. Afinal, como é possível garantir que o item foi respondido corretamente por acaso? Quais informações empíricas sustentam a hipótese que para cada X sujeitos que respondem incorretamente o item, Y sujeitos que o responderam corretamente foi ao acaso? Por falta de comprovação de tal suposição, este parâmetro tem sido pouco utilizado no âmbito de testes de desempenho.

3.1.2 Validade dos testes

Uma das grandes preocupações em relação aos instrumentos de avaliação é quanto a sua validade. É necessário que o teste seja adequado para o que se pretende medir. Um teste

é válido quando mede aquilo que pretende medir. Mas ressalta Vianna (1976) que tal definição é demasiadamente genérica e não consegue explicar a complexidade do conceito. Segundo o autor, é necessário, por exemplo, especificar em qual situação, para qual finalidade e para qual público o teste é válido. Mesmo assim, a questão da validade é necessária e generalizável a todos os testes, pois este deve realizar satisfatoriamente a medida para o qual foi construído (REQUENA, 1990).

Mehrens e Lehmann (1978) referem-se à validade como o grau em que um teste é capaz de atingir determinados objetivos. Objetivos esses de predição sobre o sujeito, quando se relaciona a um critério, e descrição do sujeito, que depende do conteúdo e do construto. Para Pasquali (2009b) esse parâmetro do teste é satisfeito quando o teste oferece uma medida congruente com a propriedade medida dos objetos.

A partir disso, pode-se tomar como exemplo a propriedade peso. Que tipo de instrumento é adequado para mensurar essa propriedade? Certamente seria necessário um instrumento que pudesse comportar a medida da densidade dos tecidos corporais, sendo a balança a mais adequada. Não poderíamos medir o peso de algo com uma régua, por exemplo. Pois as características desse instrumento não se adequa as propriedades a serem medidas, não são compatíveis. Dessa forma, a régua não é um instrumento válido para medir peso. O teste deve ainda ser capaz de discriminar sujeitos de condições diferentes. Como no exemplo, a balança deve ser capaz de identificar pessoas com diferentes pesos, embora bem sejam próximos.

Nas ciências psicossociais, principalmente na psicologia e na educação, a questão é bem mais complexa. Como obter medidas de aspectos não físicos, como o pensamento, emoção e a inteligência humana? Como garantir a validade dos testes de desempenho, tão utilizadas no âmbito educacional? Essa é uma questão relevante e deve ser considerada na elaboração de instrumentos de avaliação educacional.

Vários foram os tipos de validade apresentados: validade de conteúdo, validade preditiva, validade lógica, validade fatorial, validade externa, validade interna, validade, validade de hipótese, validade incremental, validade convergente, validade discriminante, validade generalizável, entre outros (PASQUALI, 2007).

No entanto, muitos autores (VIANNA, 1976; LINDEMAN, 1976; REQUENA, 1990; PASQUALI, 2009b) falam especificamente de três tipos de validade dos testes: validade de conteúdo, validade de critérios (preditiva e concorrente) e validade de construto. Para cada tipo de validade é utilizado um tipo de interpretação para o escore (VIANNA,

1976), respectivamente: a) como uma representação de certas capacidades do avaliado em relação aos conteúdos e comportamentos; b) como uma característica não medida diretamente pelo teste e; c) para predizer um comportamento futuro.

3.1.2.1 Validade de conteúdo

Em avaliação de desempenho, esse tipo de validade se refere à representatividade da amostra dos conteúdos e comportamentos de um determinado teste. Mais especificamente, refere-se a uma boa amostra de objetivos educacionais que foram utilizados no programa de ensino (LINDEMAN, 1976). Também é conhecido como validade curricular ou lógica (VIANNA, 1976).

Para garantir a validade do conteúdo de um teste não é suficiente analisá-lo após sua construção, é necessário planejar a laboração do conjunto de questões para que possam ser representativas dos comportamentos enfatizados durante um programa de ensino. Pasquali (2009b) propõem três fazes para esse procedimento:

1) Estabelecimento do conteúdo objeto de avaliação, com o intuito de evitar uma quantidade demasiada de itens sobre um conteúdo;
2) Definição dos objetivos que serão avaliados, proporcionando que venham garantir a multiplicidade de processos psicológicos, tais como de compreensão, aplicação, etc;
3) Determinação da proporção de cada conteúdo a ser contemplado no teste.

A partir disso, os itens do teste podem passar por uma análise do conteúdo e das características técnicas por especialistas na área, em que deverão identificar: a) os comportamentos listados devidamente representados e b) a representação das áreas de conteúdos selecionadas (VIANNA, 1976).

Como aponta Requena (1990), a validade de conteúdo não pode ser estimada adequadamente através de testes estatísticos, embora estes possam exercer papel auxiliar. Para a autora, a análise semântica do teste por um avaliador especialista na área tem papel fundamental nessa validade.

3.1.2.2 Validade de critério

Um teste apresenta validade de critério quando se mostra eficaz em predizer um desempenho específico de um sujeito, em que tem o próprio desempenho como o critério avaliado a partir de técnicas ou dados (informações) independentes do teste que se pretende validar (PASQUALI, 2009b). Indica também que o teste tem capacidade de predição de comportamento futuro dos sujeitos em situações específicas (REQUENA, 1990). Existem dois tipos de validade de critério: preditiva e concorrente.

A validade preditiva e a validade concorrente referem-se ao grau de correlação entre os escores do teste e os escores de uma medida do critério realizada independente do teste a ser validado e que mede o mesmo desempenho, o que as diferencia é o tempo em que são aplicadas, pois enquanto na primeira o teste critério foi aplicado em outro momento, na segunda o teste critério é aplicado paralelamente (VIANNA, 1976). Quando se obtém uma alta relação entre esses escores, diz-se que o teste tem validade de critério. O autor considera um valor da relação, a partir do modelo de Pearson, de no mínimo 0,45 numa escala de 0 a 1. Esse coeficiente é denominado de coeficiente de validade (LINDEMAN, 1976; REQUENA, 1990).

Este tipo de validade mostra-se necessária quando o objetivo do teste é a classificação ou seleção de pessoas (ANASTASI, 1976), como é o caso, por exemplo, das provas de desempenho escolar ou acadêmico. Isso se justifica à medida que é importante para o avaliador distinguir pessoas com diferentes graus de desempenhos e precisa de informações confiáveis para tomada de decisões importantes como aprovar ou reprovar um aluno, por exemplo.

3.1.2.3 Validade de construto

A validade de construto tem o objetivo de identificar em que medida as respostas de um teste tem um significado e determinar o grau de consistência na relação empírica do teste com esse significado (REQUENA, 1990). Segundo Mehens e Lehmann (1978) refere-se ao grau em que uma teoria ou conceito podem explicar os escores de um teste.

Para Pasquali (2009b) esse tipo de validade é a forma mais fundamental, uma vez que ela permite comprovar a capacidade de representação comportamental do teste. Nas palavras de Anastasi (1976) quando um teste apresenta esse tipo de validade, significa que ele mede o "conceito teórico" ou o traço latente a que se propunha.

Para se estabelece a validade de construto de um teste é necessário que se observe algumas recomendações, como apresenta Requena (1990): 1) Formulação de uma definição teórica do construto (conceito), especificar as manifestações e as inter-relações entre os construtos; 2) Verificar essas relações experimentalmente e; 3) Elaborar uma inferência e explicação que dê sentido a validade de construto da medição de um determinado teste.

Entre as técnicas disponíveis, a mais utilizada é a da Análise Fatorial Exploratória (AFE). Quando se utiliza essa técnica denomina-se validade fatorial. Segundo Field (2009) entre os objetivos da AFE está o de entender a estrutura de um conjunto de variáveis e construir um questionário para medir uma variável subjacente. Nesse caso específico, a AFE permitirá compreender a estrutura do conjunto de itens de um teste, ou seja, se eles estão relacionados aos construtos (conceitos) a qual o teste se propõe medir.

3.1.3 Fidedignidade dos Testes

Fidedignidade de um teste refere-se a sua precisão, ou seja, ao grau com que a medida é realizada com o mínimo de erro possível repetidas vezes com os mesmo sujeitos, produzindo resultados idênticos. Segundo Vianna (1976) fidedignidade de um teste pode ser definido como o grau de estabilidade dos resultados, ou seja, a consistência interna dos escores, em que aplicando o instrumento diversas vezes nos mesmo sujeitos se produz os mesmo resultados. Uma questão importante na fidedignidade do teste é o erro, como aponta Maroco e Garcia-Marques (2006), pois quanto mais uma medida é ausente de erro, mais consistente ela é, portanto, mais confiável.

Muitos são os termos utilizados ao se referir a esse parâmetro, tais como precisão e confiabilidade. Há também nomenclaturas utilizadas quando técnicas específicas são empregadas nesse tipo de análise, como consistência interna e estabilidade (PASQUALI, 2009b).

Usualmente, na TCT, o coeficiente de fidedignidade de um teste é obtido através da correlação dos escores de testes paralelos (REQUENA, 1990; PASQUALI, 2009b). Existem várias formas de se obter a fidedignidade de um teste: teste-reteste, método da metade, fórmula de Spearman-Brawn, coeficiente Alpha de Cronbach. Será detalhado especialmente esse último pela maior aceitação dos pesquisadores e uso na validação de testes na área de psicologia e educação. Sua fórmula é dada da seguinte maneira (FIELD, 2009):

$$\alpha = \frac{N^2\,\overline{Cov}}{\sum S^2_{Item} + \sum Cov_{Item}}$$

em que,

α = coeficiente Alpha de Cronbach;

N^2 = número de itens ao quadrado;

$\overline{Cov}$ = média da covariância entre itens;

$\sum S^2_{Item}$ = soma das variâncias do item;

$\sum Cov_{Item}$ = soma das covariâncias do item.

Segundo Maroco e Garcia-Marques (2006) o Alpha de Cronbach indica o quanto que os itens contribuem de forma uniforme para a soma não ponderada de um instrumento de medida, em que apresenta valores variando de 0 a 1, sendo adequando escores acima de 0,70. Como apresenta os autores, o teste faz a suposição de que o instrumento é unidimensional, ou seja, que mede o mesmo construto. Assim, quanto maior for a correlação entre os itens maior será a homogeneidade e a consistência com que mede o mesmo construto. Afirmam também que quanto menor for a variabilidade de um item em uma amostra de sujeitos, menor será o erro de medida, portanto, maior será a consistência interna, a fidedignidade.

3.1.4 Vieses do item

Quando um item apresenta viés, entende-se que pessoas em diferentes condições (culturais, sociais, econômicas, estruturais, etc.) terão probabilidades diferenciadas em obter êxito no referido item. Podemos tomar como exemplo alunos de escola pública ou privada, de baixa e alta condição econômica, da zona urbana ou rural, que sejam de regiões ou estados distintos.

A análise dos vieses dos itens a partir da TCT é realizada através da técnica de Angoff (Pasquali, 2009b), em que "consiste em transformar as percentagens de acertos nas duas populações em valores delta e plotá-los em coordenadas cartesianas" (p. 146). Assim, se

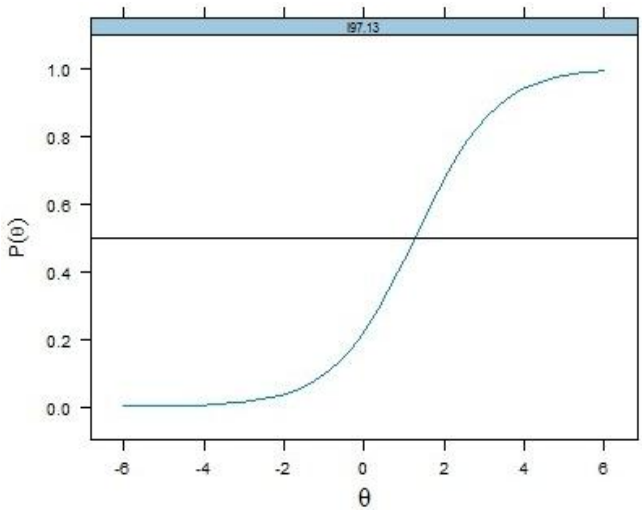
o item que não apresenta viés, altas correlações serão obtidas entre as respostas de ambos os grupos.

3.2 Teoria de Resposta ao Item (TRI)

A TRI passou a ser estruturada tecnicamente a partir dos trabalhos de Lord na década de 1950 nos Estados Unidos e de Rasch na década de 1960 na Dinamarca (PASQUALI, 2009b), tendo contribuições importantes de outros autores como Birnbaum, Hemblenton e Swaminathan (MUÑIZ, 1990). Também tem sido denominada de "Teoria do Traço Latente", "Modelos de Traço Latente", "Modelos Estruturais Latentes", "Teoria da Curva Característica do Item", mas convencionou-se, a partir de Lord a ser denominada de "Teoria de Resposta ao Item" (REQUENA, 1990).

Esse modelo trabalha, diferentemente da TCT, com o traço latente, ao nível do teta (θ), fenômeno psíquico, como critério a partir dos itens do teste (variáveis observáveis), assim, a qualidade do teste é determinada em função dos itens, com isso, ela objetiva construir itens de qualidade (PASQUALI, 2009b). A TRI sugere que a probabilidade de acertar o item decorra dos seus parâmetros e do traço latente (ou aptidão) do indivíduo exigida em um teste (VALLE, 2000). Dessa forma, esse modelo apresenta uma vantagem em relação ao clássico, a possibilidade de comparabilidade dos resultados de indivíduos de uma mesma população, mesmo que submetidos a itens diferentes (PASQUALI, 2009b; VALLE, 2000; ANDRADE, TAVARES; VALLE, 2000; MUÑIZ, 1990). A TRI busca compreender quais fatores afetam a probabilidade de um item ser acertado ou errado ou de ser aceito ou rejeitado. Segundo Pasquali (2009), para essa teoria, a probabilidade de acerto a um item aumenta em indivíduos que apresentam maior aptidão e vice-versa. Essa probabilidade é representada pelo que se denomina de Curva Característica do Item (CCI), como mostra a Figura 2.

Figura 2 – Curva Característica do Item.



Fonte: Elaboração do autor.

Como se pode observar no gráfico, na curva em formato de S ascendente, que é a característica do item, a probabilidade de acertar o item (P θ, na ordenada) aumenta em indivíduos com maior aptidão (θ, na abscissa). Essa curva pode ser afetada por vários parâmetros dependendo do modelo utilizado.

3.2.1 Parâmetros da TRI: dificuldade, discriminação e acerto casual

O primeiro parâmetro a ser considerado no item, talvez o mais importante, é o da dificuldade, em que é representado pela letra *b*. Diferentemente da TCT, a TRI o considera na mesma escala do traço latente, ou seja, do teta (θ). Dessa forma, a dificuldade está relacionada ao nível do teta necessário para responder o item (LAROS, 2009; MARTÍNEZ ARIAS; LLOREDA; LLOREDA, 2006). Na CCI seu valor é dado em termos de teta no momento em que a curva atinge 50% de probabilidade de acerto. Por exemplo, se um determinado item tem θ igual a 2, o sujeito deverá ter uma aptidão equivalente para ter 50% de chances de responder corretamente ao item. Por outro lado, se o sujeito tiver um teta inferior ou superior, o item será difícil ou fácil, diminuindo ou aumentando a probabilidade de ser acertado, respectivamente. Assim, quanto maior a dificuldade do item, maior será aptidão exigida do sujeito.

O índice de discriminação também é um parâmetro a ser considerado nos itens. Ele informa a capacidade deste em distinguir sujeitos com habilidades (aptidão) distintas,

sendo representado pela letra *a*. Assim, quanto mais o item consegue diferenciar sujeitos com magnitudes próximas de habilidade, mais discriminativo ele é (LAROS, 2009). Itens com boa discriminação permite diferenciar sujeitos que têm aptidão inferior ou superior a dificuldade do item (MARTÍNEZ ARIAS; LLOREDA; LLOREDA, 2006). Na CCI ele é dado pelo grau de inclinação no momento da inflexão ao atingir 50% de chances do item ser acertado. Quanto maior for sua inclinação, maior será a discriminação (MUÑIZ, 1990).

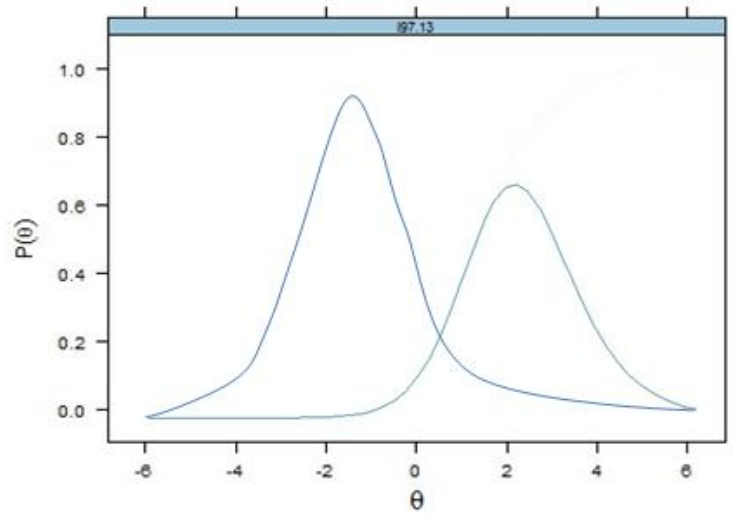
Outro parâmetro do item considerado é o acerto ao acaso. Refere-se à probabilidade de acertar o item quando não se tem aptidão suficiente (MUÑIZ, 1990) e é representado pela letra *c*. Como este parâmetro se dá em termos de proporção, ele pode variar de 0 a 1. Quanto maior o valor, maior a probabilidade dos indivíduos acertarem o item ao acaso, dado uma aptidão inferior a dificuldade do item. Esse parâmetro também está relacionado à discriminação do item. Quanto mais inclinada for a curva, mais discriminativo o item é, e menor é a probabilidade de acerto casual.

3.2.2 Função de Informação do Item

Função de Informação do Item possibilita analisar o quanto que um item oferece de informação sobre a medida da habilidade dos sujeitos. Mas também, e talvez mais importante, as informações para determinado nível do teta (MUÑIZ, 1990). Segundo Requena (1990) a informação sobre o item:

a) Varia de acordo com a aptidão, sendo diferente em cada ponto da escala;
b) Depende da CCI. Quanto maior a ascendência da curva, maior a informação;
c) Depende da variância das pontuações. Quanto maior a variância, menor a informação;
d) Para os modelos logísticos de um parâmetro, quando o teta é igual a dificuldade do item, maior é a informação.

Figura 3 – Curva de informação do Item.



Fonte: Elaboração do autor.

A função de informação do item é muito utilizada pelos construtores de testes, pois permite combinar itens de acordo com as suas necessidades, podendo diminuir o número de itens de um teste, selecionando apenas os que apresentem informações sobre a medição (MUÑIZ, 1990). Por exemplo, uma instituição pretende selecionar sujeitos da mais alta competência, levando em consideração que existem poucas vagas e muita concorrência, deverá se escolher itens que proporcione o máximo de informações para um nível alto de teta.

3.2.3 Pressupostos da TRI: unidimensionalidade e independência local

Ao analisar um conjunto de itens a partir da TRI, estes devem dispor de algumas características para que ocorra um adequado ajuste do modelo pretendido. São dois, a unidimensionalidade e independência local. Supõe-se que ao aplicar um teste, ou seja, um conjunto de itens, a probabilidade de acertá-los dependerá unicamente do traço latente do sujeito, do seu θ (MUÑIZ, 1990). Dessa forma, pressupõe-se que os itens estejam medindo um único traço latente, ou seja, que sejam unidimensionais.

Para Pasquali (2009b) apesar de ser consenso entre os psicólogos que qualquer desempenho humano seja multideterminado, em que mais de um fator latente seja responsável por uma atividade, o postulado da unidimensionalidade é suficiente, pois se admite existir um fator dominante que dê conta do conjunto de itens, a qual se supõe está sendo medido pelo teste. Nas palavras de Andrade, Tavares e Valle (2000) a resolução de um conjunto de itens deve ser em função de apenas uma habilidade.

O método mais utilizado para a comprovação desse pressuposto é a Análise Fatorial, realizada a partir de uma matriz de correlação tetracórica (ANDRADE; TAVARES; VALLE, 2000; MUÑIZ, 1990). No entanto, raras às vezes se encontra unidimensionalidade perfeita, ou seja que apenas um fator explique toda a variância, convertendo-se, dessa forma, em uma questão de grau, em que quanto mais variância for explicada pelo primeiro fator, mais unidimensionalidade existirá no conjunto de itens (MUÑIZ, 1990).

Outro pressuposto da TRI é o da independência local. Supõe-se que as respostas a um determinado item não seja influenciado por outros itens. Se existe independência local, a probabilidade de acerta um conjunto de itens é igual ao produto da probabilidade de acerta cada um destes (PASQUALI, 2009b; MUÑIZ, 1990). Com isso, o fato de um sujeito dever responder um item antes que outro, ou seja, que a resposta correta de um torne-se condição para outro, implica que os itens não sejam independentes (REQUENA, 1990). Segundo Requena (1990) e Muñiz (1990) independência local também se refere aos examinados, ou seja, se há influência de qualquer ordem entre suas respostas aos itens, suas pontuações, seus rendimentos não são independentes.

Para Muñiz (1990), se se cumpre a unidimensionalidade, matematicamente deve existir independência local, uma vez que a variável unidimensional da conta da variância dos outros itens, pois, ao contrário, seria uma contradição. Assim, tem-se apenas um pressuposto e não dois a serem verificados, uma vez que um está intrínseco no outro, devendo, dessa forma, os itens serem elaborados para atender a unidimensionalidade para que haja adequação dos modelos (ANDRADE; TAVARES; VALLE, 2000).

Uma questão importante é levantada por Tavares (2013). A autora coloca em dúvida a garantia desses pressupostos, fazendo as seguintes perguntas: será mesmo possível resguardar a unidimensionalidade, uma vez que todo fenômeno humano é multifacetado, ou seja, envolve diversos domínios? Será mesmo possível conseguir independência local nas avaliações em que são utilizados grande número itens, como é o caso do ENEM? Questões como essas merecem reflexão, uma vez que colocam em jogo a validade de um método tão aceito atualmente.

3.2.4 Modelos logísticos da TRI

Existem vários modelos de TRI na literatura e como coloca Valle (2000), estes dependem fundamentalmente de três fatores: a natureza dos itens (dicotômicos ou não), o

número de populações envolvidas e a quantidade de traços latentes medidos (quando mede mais de um denominam-se modelos multidimensionais). Ressalta a autora que os modelos unidimensionais para itens dicotômicos são os mais utilizados. Com isso, a eles será dada ênfase. Os modelos unidimensionais são classificados de acordo com os parâmetros utilizados (dificuldade, discriminação e probabilidade de acerto ao acaso), como é apresentado a seguir (PASQUALI, 2009b; VALLE, 2000; MUÑIZ, 1990):

Modelo logístico de um parâmetro: dificuldade;
Modelo logístico de dois parâmetros: dificuldade e discriminação;
Modelo logístico de três parâmetros: dificuldade, discriminação e probabilidade de acerto ao acaso.

3.2.4.1 Modelo logístico de 1 parâmetro

Esse modelo utiliza apenas o parâmetro de dificuldade dos itens, ou seja, a resposta ao item depende apenas deste e da aptidão do indivíduo, ou seja, da variável latente (ABAD et al., 2006; MUÑIZ, 1990). Foi inicialmente desenvolvido por Rasch, em que era apresentado em formato de ogiva normal, sendo adaptado para o modelo logístico por Wright (PASQUALI, 2009b; MUÑIZ, 1990). Esse modelo possibilita um tratamento matemático mais fácil (PASQUALI, 2009b) e é um dos mais populares, uma vez que apresenta uma simplicidade lógica (MUÑIZ, 1990). A expressão matemática utilizada para esse modelo é:

$$Pi\ (\theta) = \frac{e^{D(\theta - b_i)}}{1 + e^{D(\theta - b_i)}}$$

em que,

Pi (θ) = é a probabilidade de um sujeito acertar o item tendo um determinado teta;

θ = é o nível de habilidade do sujeito;

i = é o número do item no teste;

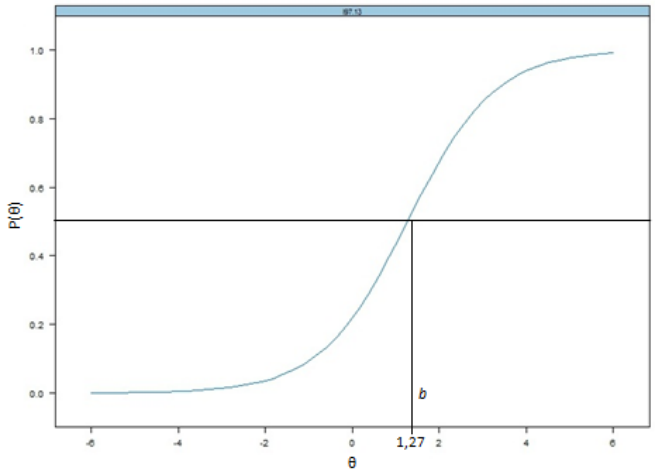
b = é o índice de dificuldade do item;

e = é uma base logarítima de valor 2,72;

D = é uma constante de valor 1,7.

Como exemplo, foi estimado o parâmetro de dificuldade para três itens de Educação Física do ENEM a partir das respostas dos candidatos de 2013. Para tanto os itens foram dicotomizados em 0 e 1, em que 0 corresponde as respostas erradas e 1 as certas. Com isso, obtemos o valor de *b*=1,27 para o índice de dificuldade do item. O modelo é representado na Figura 4.

Figura 4 – Modelo Logístico de 1 Parâmetro.



Fonte: Elaboração do autor.

Aplicando na equação, um sujeito hipotético com teta 2, terá 68% de chance de acerta esse item, como se segue:

$$Pi\ (\theta = 2) = \frac{2{,}72^{1{,}7(2-1{,}27)}}{1 + 2{,}72^{1{,}7(2-1{,}27)}} = \frac{2{,}72^{1{,}24}}{3{,}72^{1{,}24}} = \frac{3{,}46}{5{,}09} = 0{,}68$$

Para Requena (1990) na prática esse modelo apresenta vantagens e inconvenientes. O ponto positivo é que ele não necessita de grandes amostras para seu ajuste, ao contrário dos demais modelos. No entanto, por supor hipóteses muito fortes, como a que todos os itens têm o mesmo poder discriminante, é muito mais difícil se ajustar aos dados reais.

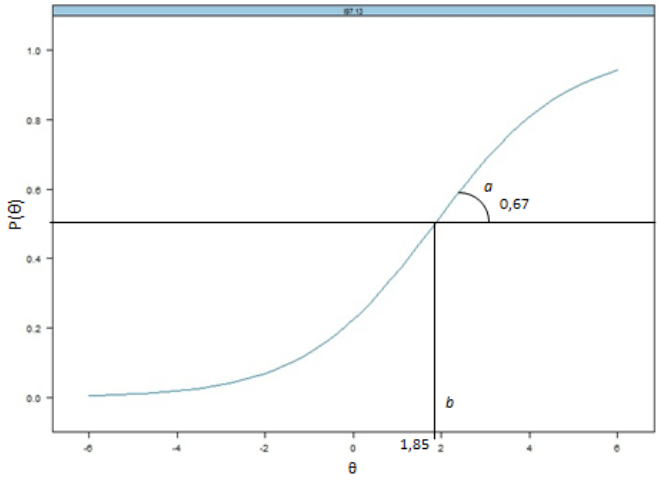
3.2.4.2 Modelo logístico de 2 parâmetros

Esse modelo considera dois parâmetros, a dificuldade e a discriminação do item. Foi inicialmente desenvolvido por Birnbaum, em que a função logística deve considerar estes dois parâmetros (PASQUALI, 2009b; MUÑIZ, 1990). A fórmula para esse modelo acrescenta-se o *a*, que é o índice de discriminação do item, dada da seguinte forma (MUÑIZ, 1990):

$$Pi\ (\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}$$

Da mesma forma, foram estimados os dois parâmetros para os mesmo itens utilizados no modelo anterior, em que obteve-se os seguintes valores: a=0,67 e b=1,85. O modelo é representado na Figura 5.

Figura 5 – Modelo Logístico de 2 Parâmetros.



Fonte: Elaboração do autor.

Tomando os parâmetros do item, um sujeito que tenha um teta de 2, terá uma probabilidade de 94% de acerta o referido item, como exemplificado a seguir:

$$Pi\ (\theta = 2) = \frac{2{,}72^{1{,}7\times0{,}67\ (2-1{,}85)}}{1 + 2{,}72^{1{,}7\times0{,}67\ (2-1{,}85)}} = \frac{2{,}72^{0{,}17}}{3{,}72^{0{,}17}} = \frac{1{,}18}{1{,}25} = 0{,}94$$

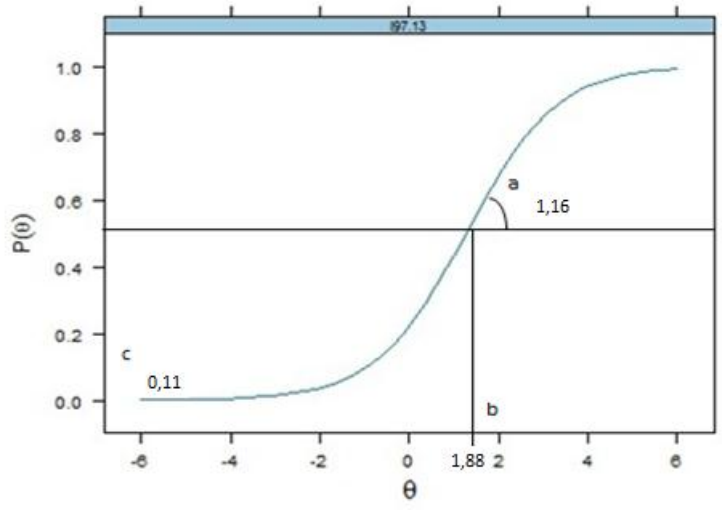
3.2.4.3 Modelo logístico de 3 parâmetros

Com origens em Birnbaum e desenvolvida por Lord, esse modelo considera três parâmetros, a dificuldade, discriminação e a probabilidade de acerto ao acaso (PASQUALI, 2009b; MUÑIZ, 1990). Para Andrade, Tavares e Valle (2000) esse modelo é o mais utilizado atualmente. A equação dada para esse modelo acrescenta-se o parâmetro *c*, que é a probabilidade de acerto ao acaso, apresentado a seguir (MUÑIZ, 1990):

$$Pi\ (\theta) = c_i + (1 - c_i)\ \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}$$

Os parâmetros desse modelo também foram estimados para os mesmo itens utilizados nos anteriores, obtendo-se os seguintes valores: *a*=1,16, *b*=1,88 e *c*=0,11. O modelo é representado na Figura 6.

Figura 6 – Modelo Logístico de 3 Parâmetros.



Fonte: Elaboração do autor.

Um determinado sujeito que tenha um teta de 2, terá uma probabilidade de 0,92% de acerta o referido item, como mostrado na equação a seguir:

$$Pi\ (\theta = 2) = 0{,}11 + (1 - 0{,}11)\frac{2{,}72^{1{,}7\times1{,}16(2-1{,}88)}}{1 + 2{,}72^{1{,}7\times1{,}16(2-1{,}88)}} = \frac{2{,}72^{0{,}24}}{3{,}72^{0{,}24}} = \frac{1{,}27}{1{,}37}$$

$$= 0{,}92$$

3.2.5 Vieses do item para a TRI

Por vezes, os itens de um teste de rendimento podem apresentar probabilidades de êxito diferenciadas entre grupos distintos. Tais efeitos podem decorrer das diferentes características e contextos culturais entre os participantes de um mesmo teste. Quando um item apresenta essas condições, ele pode oferecer uma medida errada do conhecimento em grupos heterogêneos.

Esse tipo de análise do item é denominada de Funcionamento Diferencial do Item (DIF, sigla do termo inglês). Segundo Andriola (2006, p. 117) a análise do DIF pretende responder a seguinte pergunta: “itens de testes padronizados têm o mesmo comportamento estatístico para diferentes subgrupos de sujeitos extraídos da mesma população?”.

Um item apresenta DIF quando sujeitos com habilidades equivalentes para um determinado construto, possuem diferentes probabilidades de acertar ou escolher determinada alternativa do item que mede a aptidão avaliada (ANDRIOLA, 2001). De acordo com Andriola (2001) existe uma lógica subjacente à detecção do DIF em itens, as quais seguem os seguintes procedimentos: i) estimar os parâmetros dos itens para os grupos de interesse; ii) adequar os parâmetros na mesma escala; iii) gerar a CCI dos itens para cada grupo; iv) compará-las; analisar a discrepância entre as duas curvas a partir de dados estatísticos.

Entre as técnicas para a identificação de DIF em um item está a de Mantel-Haenszel – MH (SISTO, 2006; FIDALGO; ESCALON, 2012). Explica Fidalgo e Escalon (2012) que “A lógica por trás do procedimento MH é a seguinte: se o item não apresenta DIF, a razão entre o número de pessoas que acertam o item e aquelas que o erram deve ser a mesma nos dois grupos comparados em todos os níveis de pontuação”. Os autores propõem o software GMHDIF para a análise de DIF a partir desse método.

4 PROCEDIMENTOS METODOLÓGICOS

4.1 População e amostra

Os microdados dos resultados das provas de cada candidato do ENEM estão disponíveis no site INEP em tabelas do SPSS (*Statistical Package for Social Sciences*) e são de livre acesso à comunidade acadêmica. Neste estudo utilizaram-se os microdados disponíveis a partir de 2009, quando ocorreu mudança na metodologia de análise do desempenho dos candidatos.

A população é constituída pelos candidatos cearenses participantes do ENEM entre os anos de 2009 e 2014. Para este estudo, selecionaram-se os candidatos por amostragem aleatória simples. O tamanho da amostra foi estimado a partir da seguinte equação:

$$n = \left(\frac{Z^2_{\alpha/2}\ 0{,}25}{E^2} \right)$$

Em que,

n = é o número de indivíduos na amostra;
Z α/2 = é o valor crítico que corresponde ao grau de confiança desejado;
α = nível de significância;
E = é a margem de erro ou Erro Máximo de Estimativa ($\bar{X}$ - μ). Foi utilizado um erro de 5 % (0,05).

A partir disso foi obtido um tamanho amostral de 385 candidatos para todos os anos. No caso dos anos de 2009 e 2014 optou-se por inserir mais casos a fim de se alcançar a normalidade dos mesmos. Para estimar a normalidade dos dados utilizou-se o teste Kolmogorov-Smirnov (K-S). Foram considerados normais os dados que apresentaram valor de $p > 0{,}05$. Adotou-se como variável principal para o teste de normalidade a nota em Linguagens e Códigos, área que os itens de Educação Física estão contemplados. Com isso, as amostras apresentaram normalidade com valores de *K-S* variando entre 0,02 a 0,04 e *p* entre 0,08 a 0,20. A Tabela 1 apresenta as características das amostras de cada ano.

Tabela 1 – Caracterização da amostra.

Variáveis		**2009**		**2010**		**2011**		**2012**		**2013**		**2014**	
Idade (média e desvio padrão)		19,03 (2,12)		18,99 (4,80)		18,60 (3,96)		18,66 (3,94)		18,63 (4,15)		18,91 (4,49)	
		n	%	n	%	n	%	n	%	n	%	n	%
Sexo	Fem.	237	61,4	238	61,8	216	56,1	220	57,1	215	55,8	213	54,6
	Masc.	149	38,6	147	38,2	169	43,9	165	42,9	170	44,2	177	45,4
	Total	386	100	385	100	385	100	385	100	385	100	390	100
Tipo de instituição	Pública	310	80,3	327	84,9	338	87,8	349	90,6	339	88,1	363	93,1
	Privada	76	19,7	58	15,1	47	12,2	36	9,4	46	11,9	27	6,9
	Total	386	100	385	100	385	100	385	100	385	100	390	100
Dependência Administrativa	Municipal	3	0,8	1	0,3	X	X	1	0,3	X	X	3	0,8
	Estadual	302	78,2	324	84,2	336	87,3	346	89,9	337	87,5	356	91,3
	Federal	5	1,3	2	0,5	2	0,5	2	0,5	2	0,5	X	X
	Privada	76	19,7	58	15,1	47	12,2	36	9,4	46	11,9	31	7,9
	Total	386	100	385	100	385	100	385	100	385	100	390	100
Tipo de ensino	Ensino regular	372	96,6	343	89,1	370	96,1	373	96,9	364	94,5	365	93,6
	EJA	5	1,3	16	4,2	13	3,4	10	2,6	21	5,5	23	5,9
	Ed. Profis	7	1,8	25	6,5	X	X	X	X	X	X	X	X
	Educação especial	1	0,3	1	0,3	2	0,5	2	0,5	2	0,5	2	0,5
	Total	386	100	385	100	385	100	385	100	385	100	390	100
Localização	Rural	6	1,6	3	0,8	11	2,9	20	5,2	112	3,1	9	2,3
	Urbana	380	98,4	382	99,2	374	97,1	365	94,8	373	96,9	381	97,7
	Total	386	100	385	100	385	100	385	100	385	100	390	100
Situação de Conclusão	Concluí o EM	1	0,3	1	0,3	X	X	X	X	X	X	X	X
	Concluirei o EM neste ano	371	96,1	384	99,7	384	99,7	385	100	385	100	390	100
	Concluirei o EM após este ano	14	3,6	X	X	1	0,3	X	X	X	X	X	X
	Não concluí e não estou cursando o EM	X	X	X	X	X	X	X	X	X	X	X	X
	Total	386	100	385	100	385	100	385	100	385	100	390	100
Cor/raça	Não declarado	X	X	9	2,3	10	2,6	4	1,0	6	1,6	6	1,5
	Branca	X	X	107	27,8	97	25,2	90	23,4	63	16,4	66	16,9
	Preta	X	X	21	5,5	32	8,3	28	7,3	24	6,2	31	7,9
	Parda	X	X	244	63,4	233	60,5	254	66,0	268	69,6	277	71,0
	Amarela	X	X	3	0,8	10	2,6	6	1,6	17	4,4	8	2,1
	Indígena	X	X	1	0,3	3	0,8	3	0,8	7	1,8	2	0,5
	Total	X	X	385	100	385	100	385	100	385	100	390	100

Tem necessidades especiais	Sim	1	0,3	1	0,3	2	0,5	4	1,0	6	1,6	4	1,0
	Não	385	99,7	384	99,7	383	99,5	381	99,0	379	98,4	386	99,0
	Total	386	100	385	100	385	100	385	100	385	100	390	100

Fonte: Da pesquisa.

Foram excluídos os candidatos que não indicaram o sexo, tipo de ensino, situação de conclusão do Ensino Médio, dependência administrativa da escola, localização da escola (zona rural ou urbana), cor/raça e os que não estiveram presentes na prova de Linguagens e Códigos. Também se optou por selecionar os candidatos que responderam o caderno de prova de cor azul.

4.2 Delineamento da pesquisa

A presente pesquisa se enquadra no viés quantitativo, pautada nos pressupostos da Teoria da Mensuração (PASQUALI, 2009). Caracteriza-se como descritiva e exploratória, uma vez que, procura descrever um instrumento de avaliação (o ENEM) através da exploração de sua estrutura estatística. Foram analisados os itens da prova de Linguagens e Códigos do ENEM, e de forma mais específica os parâmetros dos itens de Educação Física do exame.

4.3 Análise dos dados

Para a análise dos dados, de início foi utilizado os recursos da Análise Fatorial Exploratória pelo método da Fatoração dos Componentes Principais para verificar a unidimensionalidade da prova de Linguagens e Códigos, a qual contém os itens de Educação Física. Para tanto, analisou-se primeiro a adequação da amostra de variáveis (itens) através do teste de Kaiser-Mayer-Olklin (KMO), esfericidade de Bartlett (BTS). Considerou-se como adequado os valores de KMO $\geq 0,70$ e BTS $\leq 0,05$. Além disso, foi analisado a adequação de cada item através da Correlação de Pearson (r), sendo considerado adequado o item que apresentou $r \geq 0,50$. Em seguida solicitou-se a variância acumulada considerando um fator, uma vez que a prova é elaborada de forma a constituir uma dimensão. Também utilizou-se como parâmetro de análise da estrutura fatorial da prova o gráfico Scree Plot. Na análise da adequação de cada item ao fator considerou-se itens com cumunalidade e cargas fatoriais acima de 0,4.

Também foi realizado o cálculo de consistência interna (α de Cronbach) para a análise da fidedignidade da prova e o T^2 de Hotelling (T^2) para a existência de efeito halo (PASQUALI, 2009b; HAIR *et al.*, 2005; FIELD, 2009). Foram considerados adequados os valores $\alpha \geq 0,7$ e T^2 com $p \leq 0,05$. Além disso, foi estimado a sensibilidade do teste, sendo considerando mais sensíveis os valores mais próximos de 1.

Em seguida foram analisados os parâmetros de dificuldade e discriminação dos itens, sendo este último realizado através da correlação bisserial por ponto ($r_{bp)}$, uma vez que os itens foram dicotomizados em certo e errado (PASQUALI, 2009b). Os valores de dificuldade foram classificados da seguinte forma: muito fácil de 0 a 0,2; fácil de 0,2 a 0,4; médio de 0,4 a 0,6; difícil de 0,6 a 0,8 e; muito difícil de 0,8 a 1,0. Para o segundo parâmetro, foram considerados discriminativos os itens com $r_{bp} \geq 0,20$. As seguintes análises foram conduzidas com o software SPSS, versão 20.0.

5 RESULTADOS

5.1 Análise da prova: Análise Fatorial Exploratória, Consistência Interna das Provas e Sensibilidade

Inicialmente foram analisados os valores de correlação dos itens de Linguagens e Códigos, com destaque para os itens de Educação Física através de uma matriz de anti-imagem para verificar sua adequabilidade para compor a prova. Posteriormente verificados os valores de comunalidade e as cargas fatoriais dos itens considerando que o instrumento apresenta um fator. A correlação apresentou adequabilidade de todos os itens de Educação Física, com valores de r variando entre 0,54 a 0,89. Embora, alguns itens apresentaram valores abaixo de 0,50, o que inviabilizaria a continuação da análise fatorial, sendo necessária a exclusão de alguns itens. As comunalidades dos itens foram baixas, variando entre 0,04 e 0,30 e cargas fatoriais entre 0,20 e 0,54, com exceção do item 97 de 2014. Os valores referentes aos itens de Educação Física estão dispostos na Tabela 2.

Tabela 2 – Valores de adequação dos itens.

Ano	Item	r*	Comunalidade	Cargas Fatoriais
2009	103	0,76	0,13	0,35
	134	0,54	0,04	0,20
2010	106	0,81	0,12	0,35
	110	0,88	0,30	0,54
	120	0,69	0,07	0,26
2011	96	0,89	0,22	0,47
	105	0,87	0,30	0,39
	108	0,79	0,08	0,28
	133	0,89	0,23	0,47
2012	96	0,73	0,11	0,33
	100	0,61	0,04	0,20
	115	0,74	0,11	0,34
2013	97	0,72	0,07	0,26
	108	0,83	0,21	0,46
	111	0,78	0,11	0,32

2014	97	0,54	0,01	0,08
	98	0,72	0,18	0,42

*Correlação na matriz de anti-imagem.
Fonte: Da pesquisa.

Na Tabela 3 estão os valores referentes à adequação da prova de Linguagens e Códigos para a realização da análise fatorial. São apresentados os valores de KMO, testes de esfericidade de Bartlett e a variância explicada com um fator. O teste de KMO apresentaram valores adequados, com exceção das provas de 2012 e 2014, mas permanecendo em níveis aceitáveis. O teste de esfericidade apresentou significância em todos os anos. Esses valores indicam a adequação da amostra de itens para a realização da análise fatorial. Ao realizar a análise solicitando a extração de um fator, obteve-se uma variância explicada muito baixa (7,80 a 14,71).

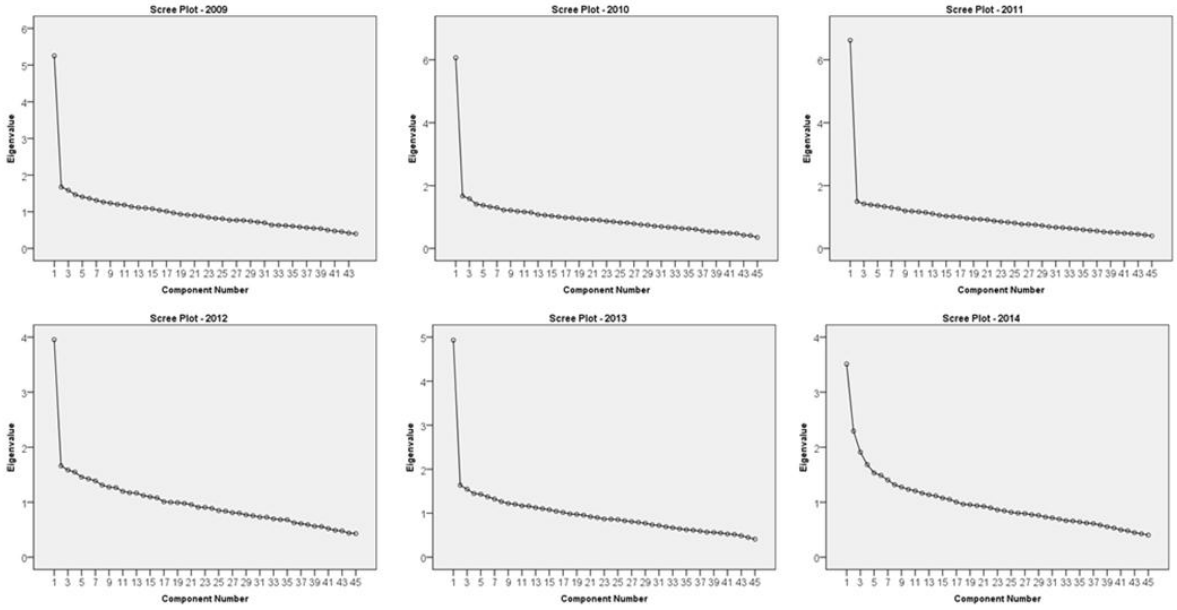
Tabela 3 – Valores de adequação do teste.

Ano	KMO	Teste de esfericidade de Barlett's			Variância explicada
		X^2	gl	Sig.	
2009	0,77	2116,96	946	0,00	11,94
2010	0,81	2431,65	990	0,00	13,48
2011	0,84	2568,24	990	0,00	14,71
2012	0,67	1654,79	990	0,00	8,79
2013	0,76	1917,90	990	0,00	10,96
2014	0,66	1859,68	990	0,00	7,80

Fonte: Da pesquisa.

A análise gráfica indica que os testes são unifatoriais, apesar da baixa variância explicada do fator das provas. A exceção foi a prova de 2014 que apresentou uma estrutura fatorial insatisfatória. A Figura 7 mostra o gráfico scree plot que apresenta a relação dos componentes e os autovalores.

Figura 7 – Scree plot das provas do ENEM de 2009 a 2014.



Fonte: Da pesquisa.

Na Tabela 4 são apresentados os valores de confiabilidade da prova. A sensibilidade das provas apresentaram valores de 0,85 a 0,99 e erro de medida entre 2,90 e 3.19. O teste T^2 de Hotteling apresentou valores satisfatórios, indicando a inexistência de efeito de halo. Com relação a precisão da prova, apenas os anos de 2009 a 2011 tiveram valores adequados. Ressalta-se a baixa precisão ($\alpha = 0,53$) da prova de 2014. De modo geral, os itens não influenciam a precisão caso sejam excluídos.

Tabela 4 – Valores de confiabilidade da prova.

Ano	Sensibilidade	EM	T^2 da prova	α	Item	Correlação Item-Total	α se o item for excluído
2009	0,98	3,19	$T^2 = 4848,22$ $F = 100,45$ $p \leq 0,05$	0,80	103	0,29	0,79
					134	0,15	0,80
2010	0,98	2,90	$T^2 = 1443,04$ $F = 29,12$ $p \leq 0,05$	0,82	106	0,28	0,82
					110	0,45	0,82
					120	0,22	0,82
2011	0,99	2,97	$T^2 = 1034,75$	0,85	96	0,40	0,85

			F = 20,88 p ≤ 0,05		105	0,33	0,85
					108	0,24	0,85
					133	0,41	0,85
2012	0,94	2,95	T² = 1101,70 F = 22,23 p ≤ 0,05	0,72	96	0,24	0,72
					100	0,17	0,72
					115	0,24	0,71
2013	0,97	2,92	T² = 1166,33 F = 23,54 p ≤ 0,05	0,79	97	0,21	0,79
					108	0,36	0,78
					111	0,25	0,79
2014	0,85	2,96	T² = 2405,62 F = 48,63 p ≤ 0,05	0,53	97	0,03	0,53
					98	0,21	0,51

Fonte: Da pesquisa.

5.3 Análise dos Itens: dificuldade e discriminação

O índice de dificuldade e discriminação dos itens estão apresentados na Tabela 5. A dificuldade dos itens variou entre 0,28 e 0,88. Quanto a discriminação, a maioria apresenta discriminação acima de 0,20, com exceção do item 100 do ano de 2012 e dos dois itens de 2014. Ambos os itens deste ano apresentaram dificuldade alta. Além disso, o item de 2014 teve discriminação inversa.

Tabela 5 – Dificuldade e discriminação dos itens.

Ano de Aplicação	Item	*d*	r_{bp}	$p\ (r_{bp})$
2009	103	0,28	0,37	0,00
	134	0,66	0,20	0,00
2010	106	0,67	0,34	0,00
	110	0,32	0,58	0,00
	120	0,54	0,28	0,00
2011	96	0,57	0,45	0,00
	105	0,49	0,38	0,00

	108	0,66	0,27	0,00
	133	0,41	0,47	0,00
2012	96	0,53	0,36	0,00
	100	0,72	0,17	0,00
	115	0,64	0,35	0,00
2013	97	0,84	0,24	0,00
	108	0,40	0,49	0,00
	111	0,39	0,34	0,00
2014	97	0,88	0,00	0,99
	98	0,78	-0,06	0,20

Fonte: Da pesquisa.

6 DISCUSSÃO

6.1 Discussão da análise da prova

A análise estatística da prova objetivou conhecer as características métricas do conjunto de itens de Linguagens e Códigos. As análises iniciais de adequação da amostra de itens de Educação Física apresentaram valores parcialmente insatisfatórios. A correlação entre os itens mostra o quanto cada item está adequado para compor a prova. Ele é realizado através da matriz de anti-imagem. Na análise os itens apresentaram valores adequados (FIELD, 2009), ou seja, nenhum item de Educação Física se mostrou inapropriado, inicialmente, para fazer parte da prova. No entanto, é necessário considerar a adequabilidade de todos os itens da prova. Nessa análise alguns itens de outras disciplinas apresentaram-se insatisfatórios, havendo a necessidade de exclusão desses itens para a continuação da análise fatorial. Considerando que o ENEM é uma avaliação que tem grandes repercussões sociais, sendo necessários resultados confiáveis, essa primeira análise mostrou deficiência no instrumento.

Mesmo com alguns itens com valores de correlação abaixo do aceitável, o teste de KMO apresentou valores aceitáveis, acima de 0,7, embora considere-se ótimo os valores acima de 0,8 e com significância no teste de esfericidade (FIELD, 2009; PONTES JUNIOR et al., 2014). Essas medidas são importantes, pois indicam a adequação do conjunto de itens para compor o instrumento e prosseguir com análise fatorial.

Após analisar a adequação da amostra de itens, procedeu-se com a análise fatorial. Ressalta-se, que nesse momento optou-se por solicitar análise com a extração de uma fator, uma vez que o Inep estima a proficiência dos candidatos via TRI, a qual tem como

pressuposto a unidimensionalidade dos itens. O objetivo foi identificar se os itens atendiam a esse critério. Nessa análise, obteve-se variância explicada muito baixa considerando um fator, todos abaixo de 15%. Pontes Junior et al. (2014) destaca a necessidade de os fatores do instrumento explicarem pelo menos 50% da variância. Hair et al. (2005) coloca que muitos autores colocam a necessidade de pelo menos 60% de explicação da variância. No caso da prova de Linguagens e Códigos do ENEM, levando em conta que a organizadora analisa os resultados pela técnica da TRI, portando assumindo a unidimensionalidade dos itens, a prova está aquém dos valores adequados para a explicação da variância. Isso se torna um problema porque o teste assume que o desempenho dos estudantes é representada por uma variável latente.

Pasquali (2009) ressalta a dificuldade de se obter a unidimensionalidade, uma vez que, os aspectos cognitivos dos seres humanos são multifacetados e multideterminados. Tavares (2013) questiona o fato de avaliações educacionais pretenderem medir um fator unidimensional, uma vez que, é consensual entre os educadores pesquisadores que o ser humano é determinado por vários fatores. Mesmo assim, Pasquali (2009) propõe ser suficiente a existência de um fator dominante para garantir esse pressuposto. Em modo complementar, Vitória, Almeida e Primi (2006) afirmam, com base em pesquisas, que a dimensionalidade (os diferentes graus) pouco afeta os parâmetros dos itens e os resultados de testes.

Além disso, ao extrair um fator, os itens apresentam variâncias comuns muito baixas, não passando de 0,30, enquanto que o mínimo aceitável é 0,40, acontecendo o mesmo com as cargas fatoriais dos itens, em que muitos não atingiram o valor mínimo (PONTES JUNIOR et al., 2014). Apesar dos valores inadequados, a análise gráfica através do *scree plot* mostra a unidimensionalidade dos dados ao observar o ponto de inflexão, com exceção dos dados de 2014 que apresentaram uma estrutura inadequada.

No que se refere à precisão da prova utilizou-se o coeficiente Alpha de Cronbach, que é uma das técnicas mais utilizadas para a análise da confiabilidade de um instrumento (CUNHA; ALMEIDA NETO; STACKFLETH, 2016). Com exceção das provas de Linguagens e Códigos de 2009 e 2011 com valores α acima de 0,80, considerados excelentes, o demais tiveram valores inferiores, embora aceitáveis. Cabe destacar a baixa precisão da prova de 2014, que teve o coeficiente α de 0,53. De modo geral, os itens de Educação Física não influenciaram na precisão do instrumento, o que é considerado bom.

A precisão ou fidedignidade de um instrumento de medida é central na sua validação. Uma boa precisão é uma garantia da confiabilidade do teste. Quando se trata de testes para fins de seleção, sua importância aumenta, já que pode influenciar nos resultados.

A fidedignidade ou precisão pode ser afetada por uma série de fatores, como afirma Vianna (1976). Segundo o autor, a precisão do teste aumenta quanto maior for o número de itens, desde que com índice de correlação suficiente com os demais itens do teste. Nesse quesito, o ENEM não poderia apresentar problemas, uma vez que que cada área do exame, no caso deste estudo a área de Linguagens e Códigos, são constituídos de 45 itens, o que é considerado um grande número. Além do mais, o exame é alvo de muitas críticas no sentido de diminuir o número de itens.

Outro fator que pode influenciar na precisão da prova é a amplitude de dificuldade dos itens (VIANNA, 1996). Quanto menor a amplitude maior a fidedignidade. Considerando apenas os itens de Educação Física de 2014, observa-se que os itens apresentam baixa dificuldade, fato também observado ao considerar os demais itens. Isso pode ter contribuído para o baixo índice alpha no exame deste ano.

6.2 Discussão da análise dos itens

A análise dos parâmetros de dificuldade e discriminação dos itens a partir da TCT objetivou observar as caraterísticas estatísticas de adequação dos itens ao realizar medidas. A dificuldade apresentou índices adequados, ou seja, entre 0,20 e 0,80, com exceção do item 97 de 2014. Quanto à discriminação, a maioria permaneceu acima de 0,20. No entanto, os itens de Educação Física da prova de 2014 apresentaram discriminação muito abaixa, ou seja, são itens ruins para avaliar, uma vez que não conseguem diferenciar os sujeitos de baixa habilidade dos de alta habilidade.

O índice de dificuldade é um parâmetro simples de um teste. Neste caso, como coloca Vianna (1976), poderia ser chamado de índice de facilidade, uma vez que se utiliza da proporção de acertos. Dessa forma, quanto maior a proporção mais fácil é o item. No caso dos itens de Educação Física analisados, a maioria apresentou índices adequados.

Vale ressaltar que esse parâmetro depende do particular conjunto de sujeitos que constituem a amostra avaliada (LAROS, 2009). Dessa forma, o autor ressalta que os sujeitos habilidosos a dificuldade vai aumenta. Do mesmo modo, a dificuldade do item será baixa se os sujeitos não forem tão habilidosos. Entretanto, destaca-se que esse problema desaparece se a amostra for representativa da população, em que qualquer amostra aleatória apresentará os mesmo índices para os itens (LAROS, 2009). No caso deste estudo em particular, acredita-se

ter superado esse problema, considerando que o pressuposto da normalidade foi atendido para a amostra de todos os anos.

Por outro lado, quanto à discriminação, os itens de 2014 apresentaram baixo índice. Esses mesmos itens apresentaram dificuldade elevada. Alguns autores (VIANNA, 1976; PASQUALI, 2009) ressaltam que itens com dificuldade alta apresentam discriminação inadequada.

Nesse estudo, utilizou-se do coeficiente de correlação item-total, realizado através da correlação bisserial por ponto, uma vez que, segundo Silveira (1983) é o melhor indicador da discriminação de um item. Vários estudos que se utilizam da TCT para analisar testes utilizam-se desse parâmetro.

No entanto, Laros (2009) considera inadequado para o cálculo da discriminação por apresentar problemas teóricos. Ressalta ser incoerente avaliar um parâmetro de um item através da correlação do item com o escore total do teste, uma vez que os outros itens ainda não foram testados. Destaca ainda a necessidade de os itens do teste apresentar unidimensonalidade para que a discriminação seja consistente. Além do mais, indica que o parâmetro é falho quando se tem itens muito fáceis ou muito difíceis.

7 CONCLUSÃO

A prova de Linguagens e Códigos apresenta problemas de dimensionalidade, considerada pressuposto fundamental para validade dos testes. Apesar de a análise gráfica indicar o cumprimento desse indicador, baixa explicação foi demostrada. Isso compromete a medida realizada pelo instrumento. Quanto à precisão ou fidedignidade da prova apenas em 2014 apresentou baixo índice. Nesses aspectos os itens de Educação Física permaneceram adequados.

No geral, os itens da área apresentaram valores adequados de dificuldade e discriminação a partir da TCT, com exceção da prova de 2014, em que apresentaram valores inadequados, o que pode ter influenciado nos valores de fidedignidade e unidimensionalidade. Esses índices indicam baixa qualidade dos itens de Educação Física.

As análises empreendidas são dificultadas pela baixa representatividade dos itens de Educação Física, já que poucos itens são contemplados no exame. Esse fato compromete também a avaliação dos conteúdos da disciplina, sendo difícil obter um diagnóstico da aprendizagem nessa área com um número limitado de questões.

Diante disso, levando em consideração que o ENEM passou a ser analisado a partir da TRI em substituição a TCT, levanta-se o seguinte questionamento: A TRI trouxe avanços nas medidas educacionais em relação a TCT? Dessa forma, indica-se a realização de pesquisas que comparem os parâmetros das provas e dos itens dos exames educacionais, de forma a possibilitar informações que forneçam indicadores que proporcionem a comparação das técnicas de análise utilizadas pelas duas teorias, de modo a testar os supostos avanços pretendidos pela TRI.

REFERÊNCIAS

ABAD, F. J. et al**. Introducción a la ppsicometría: Teoría Clásica de los Tests y Teoría de la Respuesta al Ítem**. Universidad Autonoma de Madrid, 2006.

ANASTASI, A. **Testes psicológicos: teoria e aplicação**. Tradução de Dante Moreira Leite. São Paulo: EPU, 1976.

ANDRADE, D. F.; TAVARES, H. R.; VALLE, R. C.. **Teoria de Resposta ao Item: Conceitos e Aplicações**. ABE – Associação Brasileira de Estatística, 2000.

ANDRADE, J. M.; LAROS, J. A.; GOUVEIA, V. V. O uso da teoria da resposta ao item em avaliações educacionais: diretrizes para pesquisadores. **Avaliação Psicológica**, n. 9, v. 3, p.421-435, 2010.

ANDRIOLA, W. B. Descrição dos principais métodos para detectar o Funcionamento Diferencial dos Itens (DIF). **Psicologia: Reflexão e Crítica**, v. 14, n. 3, p. 643-652, 2001.

ANDRIOLA, W. B. Estudo do viés dos itens em testes de rendimento: uma retrospectiva. **Estudos em Avaliação Educacional**, v. 17, n. 35, set./dez., 2006.

ARSLAN, Y.; ERTURAN İLKER, G.; DEMİRHAN, G. Evaluation Development Program on Pre-service Physical Education Teachers' Perceptions Related to Measurement and Evaluation. **Educational Sciences: Theory & Practice**, v. 13, n. 2, Spring, p. 1119-1124, 2013.

BELTRÃO, J. A. A educação física na escola do vestibular: as possíveis implicações do ENEM. **Movimento**, Porto Alegre, v. 20, n. 2, p. 819-840, abr./jun. de 2014.

BLOOM, B. S. et. al. **Taxonomy of Educational Objectives, the classification of educational goals – Handbook I: Cognitive Domain**. New York: Mc Kay, 1956.

BONILLO, A. Análise de los ítems. In: MENESES, J. (org.). **Psicometría**. Barcelona: EDITORIAL UOC, 2013.

BRASIL. **Portaria MEC Nº 438, de 28 de maio de 1998**. Brasília, 1998.

BRASIL. Ministério da Educação. **Parâmetros curriculares nacionais: ensino médio**. Brasília: Ministério da Educação, 2000.

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais – INEP. **Exame Nacional do Ensino Médio – ENEM: Documento Básico**. Brasília, 2002a.

BRASIL. **Livro introdutório: Documento básico: ensino fundamental e médio**. Brasília: MEC: INEP, 2002b.

BRASIL. Presidência da República. **Medida Provisória nº 213, de 10 de setembro de 2004**. Institui o Programa Universidade para Todos – PROUNI. Regula a atuação de entidades beneficentes de assistência social no ensino superior, e dá outras providências. Brasília, DF, 2004.

BRASIL. Ministério da Educação. **PDE: Plano de Desenvolvimento da Educação: SAEB: ensino médio: matrizes de referência, tópicos e descritores**. Brasília: MEC, SEB; Inep, 2008.

BRASIL. Ministério da Educação. **Portaria Normativa nº 2, de 26 de janeiro de 2010**. Institui e regulamenta o Sistema de Seleção Unificada, sistema informatizado gerenciado pelo Ministério da Educação, para seleção de candidatos a vagas em cursos de graduação disponibilizadas pelas instituições públicas de educação superior dele participantes. 18. ed. Brasília, DF, 2010a.

BRASIL. Ministério da Educação. **Portaria Normativa nº 10, de 10 de abril de 2010**. Dispõe sobre procedimentos para inscrição e contratação de financiamento estudantil a ser concedido pelo Fundo de Financiamento ao Estudante do Ensino Superior (FIES). Brasília, DF, 2010b.

BRASIL. **LDB: Lei de Diretrizes e Bases da Educação Nacional: Lei nº 9.394, de 20 de dezembro de 1996, que estabelece as diretrizes e bases da educação nacional**. 8. ed. – Brasília: Câmara dos Deputados, Edições Câmara, 2013.

BRASIL. Inep. Ministério da Educação. **Edital Nº 6, de 15 de maio de 2015 Exame Nacional do Ensino Médio – ENEM 2015**. 2015 Disponível em: <http://download.inep.gov.br/educacao_basica/enem/edital/2015/edital_enem_2015.pdf>. Acesso em: 29 jul. 2015.

CAMPBELL, D. T.; STANLEY, J. C. **Delineamentos experimentais e quase-experimentais de pesquisa**. Tradução de Renato Alberto T. Di Dio. São Paulo: EPU, 1979.

CEARÁ. Secretaria da Educação. **Boletim do Sistema de Avaliação SPAECE – 2012**. Universidade Federal de Juiz de Fora, Faculdade de Educação, CAEd, Juiz de Fora, v. 3, jan./dez., 2012.

COELHO, M. I. M. Vinte anos de avaliação da educação básica no Brasil aprendizagens e desafios. **Ensaio: Aval. Pol. Públ. Educ.**, Rio de Janeiro, v. 16, n. 59, p. 229-258, abr./jun. 2008.

CORTI, A. P. As diversas faces do ENEM: análise do perfil dos participantes (1999-2007). **Est. Aval. Educ.** [online], vol.24, n.55, pp. 198-221, 2013.

CUNHA, C. M.; ALMEIDA NETO, O. P.; STACKFLETH, R. Principais métodos de avaliação psicométrica da confiabilidade de instrumentos de medida. **Rev. Aten. Saúde**, São Caetano do Sul, v. 14, n. 49, p. 98-103, jul./set., 2016.

DARIDO, S.C. **Educação física na escola: implicações para a prática pedagógica**. 2 ed. Rio de Janeiro. Guanabara Koogan, 2011.

DEPRESBITERIS, L. Avaliação de programas e avaliação da aprendizagem. **Educação e Seleção**, n. 19, p. 5-31, 1989.

FERNANDES, A.; RODRIGUES, H. A.; NARDON, T. A. A inserção dos conteúdos de educação física no ENEM: entre a valorização do componente curricular e as contradições da democracia. **Motrivivência**, nº 40, p. 13-24, Jun., 2013.

FIDALGO, Á. M; SCALON, J. D. Uso dos Métodos Mantel-Haenszel para a Detecção do Funcionamento Diferencial dos Itens e Software Relacionado. **Psicologia: Reflexão e Crítica**, v. 25, n. 1, p. 60-68, 2012.

FIELD, A. **Descobrindo estatística usando o SPSS**. Tradução de Lorí Vialli. Porto Alegre: ArtMed, 2009.

FREITAS, D. N. T. **Avaliação da educação básica no Brasil: origens e pressupostos**. In: BAUER, A.; GATTI, B. A.; TAVARES, M. Vinte e cinco anos de avaliação de sistemas educacionais no Brasil: origens e pressupostos. Florianópolis: Insular, 2013.

GATTI, B. A. **Possibilidades e fundamentos de avaliação em larga-escala: primórdios e perspectivas contemporâneas**. In: BAUER, A.; GATTI, B. A.; TAVARES, M. R. Vinte e cinco anos de avaliação de sistemas educacionais no Brasil: origens e pressupostos. Florianópolis: Insular, 2013.

HAIR, J. F. et al. **Análise multivariada de dados**. (5ª ed). Porto Alegre, RS: Bookman, 2005.

HORTA NETO, J. L. Um olhar retrospectivo sobre a avaliação externa no Brasil: das primeiras medições em educação até o SAEB de 2005. **Revista Iberoamericana de Educación**. n. 42, v. 5, 2007.

INEP. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Exame Nacional do Ensino Médio (Enem): relatório pedagógico 2009-2010**. Brasília: O Instituto, 2013.

LAROS, J. A. **Análise gráfica de itens**. In: PASQUALI, L. Psicometria: teoria dos testes na psicologia e na educação. 3 ed. Petrópolis, RJ: Vozes, 2009.

LIMA, A. C. Ciclo de avaliação da educação básica do Ceará: principais resultados. **Est. Aval. Educ.**, São Paulo, v. 23, n. 53, p. 38-58, set/dez., 2012.

LIMA, A. M. G.; GOMES, C. A. S.; ANDRIOLA, W. B. **Sistemas nacionais de avaliação: da educação básica ao ensino superior**. In: LEITE, R. H. (org.). Diálogos em avaliação educacional. Fortaleza, Edições UFC, 2014.

LINDEMAN, R. H. **Medidas educacionais: testes objetivos e outros instrumentos de medida para a avaliação da aprendizagem**. Brasília: Editora Globo, 1976.

MAROCO, J.; GARCIA-MARQUES, T. Qual a fiabilidade do alfa de Cronbach? Questões antigas e soluções modernas? **Laboratório de Psicologia**, v. 4, n. 1, p. 65-90, 2006.

MARTÍNEZ ARIAS, M. R.; LLOREDA, M. V. H.; LLOREDA, M. J. H**. Psicometría**. Alianza Editorial, 2006.

MEHRENS, W. A.; LEHMANN, I. J. **Testes padronizados em educação**. Tradução de Renato Alberto T. Di Dio e Ricardo Pinheiro Lopes. São Paulo: EPU, 1978.

MUÑIZ, J. La medición de lo psicológico. **Psicothema**, v. 10, n. 1, p. 1-21, 1998.

MUÑIZ, J. **Teoría de repuesta a los ítems**. Madrid: Pirámede, 1990.

PASQUALI, L. Validade dos Testes Psicológicos. **Psic.: Teor. e Pesq.**, Brasília, v. 23, n. esp., p. 099-107, 2007.

PASQUALI, L. Psicometria. **Rev Esc Enferm USP**, v. 43, p. 992-999, 2009a.

PASQUALI, L. **Psicometria: teoria dos testes na psicologia e na educação**. 3 ed. Petrópolis, RJ: Vozes, 2009b.

PONTES JUNIOR, J. A. F; ALMEIDA, L. S; TROMPIERI FILHO, N. Avaliação cognitiva em larga escala dos conteúdos da Educação Física escolar. **Bordón: revista de pedagogía**, v. 66, p. 9-25, 2014.

PONTES JUNIOR, J. A. F; SOARES, E. S; TROMPIERI FILHO, N. **Expectativa discente sobre os instrumentos de avaliação na educação física escolar.** In: Livro de Actas do I Seminário Internacional, aprendizagem e Rendimento. Braga: Universidade do Minho, Centro de Investigação em Educação (CIEd), 2014.a

PONTES JUNIOR, J. A. F; SOARES, E. S; TROMPIERI FILHO, N. **Utilização das escalas de medidas na avaliação da aprendizagem na educação física escolar**. In: LEITE, Raimundo Hélio (org.). Diálogos em avaliação educacional. Fortaleza, Edições UFC, 2014.b

PONTES JUNIOR; J. A. F.; TROMPIERI FILHO, N. Avaliação do ensino-aprendizagem na Educação Física escolar. **EFDeportes.com-Revista Digital**. Buenos Aires, nº 161, 2011. Disponível em: http://www.efdeportes.com/efd161/avaliacao-na-educacao-fisica-escolar.htm. Acesso em 07/10/2014.

PONTES JUNIOR; J. A. F. et al. Análise fatorial exploratória e alpha de Cronbach: elementos iniciais na validação de instrumentos de avaliação educacional. **Educação & Linguagem**, v. 1, n. 1, p. 63-75, 2014.

POPPER, K. R. **A lógica da pesquisa científica**. São Paulo: Editora Cultrix, 1972.

REQUENA, C. S.. **Psicometria: teoria y practica em la construcción de tests**. Madrid: Ediciones Norma, 1990.

SANTOS, M. F.; MARCON, D.; TRENTIN, D. T. Inserção da educação física na área de linguagens, códigos e suas tecnologias. **Motriz**, Rio Claro, v.18 n.3, p.571-580, jul./set. 2012.

SARTES, L. M. A.; SOUZA-FORMIGONI, M. L. O. Avanços na Psicometria: Da Teoria Clássica dos Testes à Teoria de Resposta ao Item. **Psicologia: Reflexão e Crític**a, v. 26, n. 2, p. 241-250, 2013.

SILVEIRA, F. L. Considerações sobre o índice de discriminação de itens em testes educacionais. **Educação & Seleção**, n. 07, 1983.

SISTO, F. F. O funcionamento diferencial dos itens. **Psico-UFS**, v. 11, n. 1, p. 35-43, jan./jun., 2006.

SOUSA, E. M. Políticas públicas e a questão da avaliação. **Ensaio: Aval. Pol. Públ. Educ.** [online], v.01, n.02, p. 51-59, 1994.

SOUZA JÚNIOR, O. M. et al. **Educação física no ENEM: análise das questões à luz dos PCN's**. II Congresso Internacional de Educação Física, Esporte e Lazer. São Carlos, SP, 2012.

TAVARES, C. Z. Teoria de resposta ao item: uma análise crítica dos pressupostos epistemológicos. **Est. Avali. Educ.**, São Paulo, v. 24, n.54, p. 56-76, jan./abr., 2013.

THOMAS, J. R.; NELSON, S. K.; SILVERMAN, S. J. **Métodos de pesquisa em atividade física**. 6ª ed., Porto Alegre: ArtMed, 2012.

VALLE, R. C. Teoria de resposta ao item. **Est. Aval. Educ.** [online]. n. 21, p. 07-92, 2000.

VIANNA, H. M. Avaliação: considerações teóricas e posicionamentos. **Est. Aval. Educ.** [online], n.16, pp. 05-36, 1997.

VIANNA, H. M. **Avaliações nacionais em larga escala: análises e propostas**. São Paulo: DPE, 2003.

VIANNA, H. M. Provas e testes no concurso vestibular. **Educação e Seleção**, n.12, p. 47-72, 1985.

VIANNA, H. M. **Testes em educação**. São Paulo: IBRASA, 1976.

VIGGIANO, E.; MATTOS, C. O desempenho de estudantes no ENEM 2010 em diferentes regiões brasileiras. **Rev. Bras. Estud. Pedagog.** (online), Brasília, v. 94, n. 237, p. 417-438, mai./ago., 2013.

VITORIA, F.; ALMEIDA, L. S.; PRIMI, R. Unidimensionalidade em testes psicológicos: conceito, estratégias e dificuldades na sua avaliação. **Psic** [online], vol.7, n.1, p. 01-07, 2006.

WAISELFISZ, J. J. O sistema nacional de avaliação do ensino público de 1º grau. **Estudos em avaliação educacional**, São Paulo, n.4, p.65-72, 1991.