Revista Brasileira de Ensino de Física, vol. 42, e20190149 (2020)
www.scielo.br/rbef
DOI: http://dx.doi.org/10.1590/1806-9126-RBEF-2019-0149

Pesquisa em Ensino de Física

Validação da nova versão do Test of Understanding Graphs in Kinematics (TUG-K) com estudantes de ensino médio

Evaluating the validity of the new version of Test of Understanding Graphs in Kinematics (TUG-K) with high school students

R. F. F. da Cunha[*1], D. G. G. Sasaki[1]

[1]Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, Rio de Janeiro, RJ, Brasil

Recebido em 23 de Junho de 2019. Revisado em 30 de Setembro de 2019. Aceito em 20 de Outubro de 2019.

Com o intuito de verificar a compreensão de gráficos cinemática de um grupo de estudantes de ensino médio, é recomendável utilizar um instrumento que possua tanto uma validade de conteúdo por parte de especialistas, quanto uma validação estatística. Nesse sentido, foi escolhida a nova versão do Test of Understanding Graphs in Kinematics (TUG-K), proposta em 2017 por Zavala e originalmente criado por Beichner, em 1994. O TUG-K foi elaborado para medir o conhecimento de gráficos em cinemática de alunos universitários, majoritariamente. Logo, para que possa ser usado na educação básica, é preciso saber se possui validade estatística nesse contexto. Portanto, foi realizada uma análise estatística do teste, após ser aplicado em dois momentos distintos, em estudantes de ensino médio de uma escola federal do Rio de Janeiro. Os parâmetros medidos foram os mesmos utilizados por Zavala. O principal resultado deste artigo foi demonstrar a validação do TUG-K nesse grupo. Como complemento, mostrou-se que o ganho de aprendizagem normalizado de Hake desses estudantes submetidos a aulas expositivas de cinemática foi de 17%, valor esperado para o ensino tradicional. A perspectiva subjacente é difundir e incentivar o uso do TUG-K no ensino médio.
Palavras-chave: Peer Instruction, gráficos de cinemática, validação do TUG-K.

In order to verify the kinematics graphs comprehension for a group of high school students, it is recommended to use an instrument that has a content valid by experts and a statistical validation. In this sense, was chosen the updated version of Test of Understanding Graphs in Kinematics (TUG-K), proposed in 2017 by Zavala and originally created by Beichner, in 1994. The TUG-K was elaborated to measure the understanding of graphs in kinematics of university students, mostly. Therefore, for this test to be used in basic education, it is necessary to know if it has statistical validity in this context. Consequently, a statistical analysis of the test was performed, after being applied at two different moments, with upper secondary level students of a federal school in Rio de Janeiro. The measured parameters were the same used by Zavala. The main result of this article was to demonstrate that TUG-K validity in this group. As a complement, it was shown that Hake's normalized learning gain from this group of students exposed to kinematics lectures was 17%, a value that is expected to be in traditional teaching. The underlying perspective is to disseminate and encourage the use of TUG-K in high school.
Keywords: Peer Instruction, kinematics graphs, TUG-K validity.

1. Introdução

O *Test of Understanding Graphs in Kinematics* (TUG-K) [1] é o teste mais utilizado no mundo para verificar os conhecimentos de estudantes, predominantemente de nível superior, sobre gráficos em cinemática e sua interpretação.

A primeira versão desse teste foi elaborada e publicada por Robert J. Beichner, professor de Física da Universidade Estadual da Carolina do Norte, em 1994 [2]. Seu objetivo principal ao desenvolver o teste era que os resultados ajudariam a expor as dificuldades dos estudantes quanto à interpretação de gráficos em cinemática. Beichner defendia que conhecer as concepções alternativas dos estudantes sobre o tema seria de grande valia para os professores antes, durante e após as aulas. Já o objetivo secundário desse seu trabalho era propor um modelo de criação de teste de múltipla escolha fruto de uma pesquisa em ensino e que possa ser utilizado como avaliação diagnóstica, formativa e/ou somativa.

É muito frequente professores de física fazerem uso de gráficos durante as aulas, de maneira natural, como se fosse uma linguagem comum a todos. Porém os resultados das aplicações do TUG-K indicam que os estudantes, de maneira geral, possuem dificuldades em compreender essa linguagem. Ou seja, há uma distância entre o que os gráficos apresentam de informação e o que os estudantes conseguem interpretar.

Essa dificuldade de interpretação de gráficos foi alvo de diversas pesquisas na década de 80 revelando as concepções alternativas (*misconceptions*) mais comuns

*Endereço de correspondência: ricardofagundes@cp2.g12.br.

dos estudantes [3,4]. Esses trabalhos motivaram Beichner a desenvolver um teste de múltipla escolha que fosse capaz de expor e mapear essas dificuldades, a fim de melhorar o ensino de Física. Beichner classificou seis tipos mais frequentes de dificuldades que os estudantes apresentam ao interpretar os referidos gráficos: (i) Gráficos como imagem do movimento: os estudantes percebem o gráfico não como uma representação matemática do movimento, mas como uma fotografia de uma característica concreta do movimento, por exemplo a sua trajetória. (ii) Confusão entre inclinação e altura: quando solicitados sobre a inclinação de um gráfico, os estudantes tendem a ler o valor em um dos eixos (iii) Confusão de variáveis: confundir os gráfico de posição, velocidade e aceleração, sem perceber que a troca entre essas grandezas no eixo vertical altera a forma funcional do gráfico. (iv) Erro de inclinação que não passa pela origem: quando solicitados a calcular a inclinação de um segmento de reta, estudantes tendem a calcular como se a reta passasse pela origem, ou seja, dividindo o valor da ordenada pela abscissa. (v) Desconhecimento da área sob a curva: não saber o significado da área sob uma curva. (vi) Confusão entre área, inclinação e altura: quando solicitados a calcular áreas, tendem a calcular inclinação ou ler a altura em um dos eixos.

Até chegar à versão publicada em 1994 (versão 2.6), Beichner apresentou um teste para quinze professores de ensino médio e de ensino superior, que contribuíram com críticas e sugestões. Após esse crivo, o teste sofreu algumas pequenas alterações em alguns itens (questões), e foi aplicado em 895 estudantes de ensino médio e majoritariamente de ensino superior. Todos eles já haviam assistido a aulas tradicionais de cinemática.

Nessa primeira versão, o TUG-K era composto por 21 itens e apresentava 7 objetivos gerais. Vinte e três anos mais tarde, em 2017, após dois anos de pesquisa, Zavala e Beichner apresentaram uma nova versão do TUG-K [5], com 26 itens ao todo (versão 4.0). Nessa mesma publicação, os autores discutiram todas as modificações realizadas, como alteração nos distratores, nos gráficos, no enunciado ou substituições e acréscimos de itens. Já os 7 objetivos gerais basilares originais (Quadro 1) permaneceram inalterados.

Antes da publicação final, essa versão 4.0 foi testada em 471 estudantes de um curso introdutório de Física em uma universidade privada no México.

No Brasil, raríssimos foram os trabalhos publicados aplicando o TUG-K versão 2.6. Em destaque, Agrello e Garg [6], que traduziram para o português e aplicaram em 228 estudantes calouros de diversos cursos da Universidade de Brasília, obtendo resultados semelhantes aos de Beichner.

Tratando-se da recente versão 4.0, ainda não há trabalhos publicados no Brasil. Nesse sentido, os múltiplos objetivos do presente trabalho são: i) divulgar a tradução para o português da versão 4.0 do TUG-K, compatível com a linguagem usada pelos professores e livros no Brasil[1]; ii) fazer uma validação estatística quantitativa dessa versão em um público-alvo específico que são estudantes de ensino médio de uma escola da rede pública federal da cidade do Rio de Janeiro; iii) comparar os resultados estatísticos no contexto brasileiro com aqueles obtidos por Zavala, destacando semelhanças e divergências; iv) sugerir alterações no próprio instrumento de avaliação (TUG-K), identificando itens cujos resultados destoam significativamente dos parâmetros escolhidos, tais como o índice de dificuldade e/ou a correlação item-total.

1.1. Análise estatística das respostas dos alunos

O TUG-K foi aplicado para estudantes de uma escola da rede federal do município do Rio de Janeiro, em dois momentos distintos: no início do ano letivo, antes mesmo de o assunto ser abordado em sala (pré-teste) e no 3° trimestre, após o assunto ser abordado nas turmas (pós-teste). A primeira aplicação contou com a participação de 127 estudantes e a última, com 118.

Conforme exposto anteriormente no Quadro 1, os 26 itens que compõem o TUG-K estão divididos em 7 grupos, cada qual com seu objetivo específico bem definido. Nesse sentido, o Quadro 2, a seguir, funciona como um complemento do quadro anterior, trazendo consigo quais são os itens pertencentes ao mesmo grupo, bem como uma pequena descrição dos mesmos.

Primeiramente, vamos apresentar os resultados provenientes da estatística descritiva. A Tabela 1 mostra a

[1] A tradução da versão 4.0 do TUG-K foi feita pela estudante de Mestrado Mariana de Almeida Jotta Barros e seu orientador, o professor Vitor Luiz Bastos de Jesus, ambos do Instituto Federal do Rio de Janeiro (IFRJ), campus Nilópolis, em parceira com o professor Daniel Guilherme Gomes Sasaki, do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca (CEFET-RJ), campus Maracanã.

Quadro 1: Cada um dos 26 itens do TUG-K trabalha um desses 7 objetivos gerais expostos nesta tabela. Na versão mais recente do TUG-K são pelo menos três itens por objetivo e, cada um tem uma abordagem específica, podendo expor diferentes dificuldades dos estudantes.

Objetivos gerais do TUG-K
1. Dado o gráfico de posição em função do tempo, determinar a velocidade
2. Dado o gráfico de velocidade em função do tempo, determinar a aceleração
3. Dado o gráfico de velocidade em função do tempo, determinar o deslocamento
4. Dado o gráfico de aceleração em função do tempo, determinar a variação de velocidade
5. Dado um gráfico de cinemática, selecionar outro gráfico correspondente
6. Dado um gráfico de cinemática, ser capaz de descrever textualmente um movimento
7. Dada a descrição textual de um movimento, selecionar o gráfico correspondente

Quadro 2: A tabela expõe de maneira mais detalhada a composição do teste, apresentando uma pequena descrição de cada item. Pode-se identificar também quais os itens que possuem o mesmo objetivo, ou seja, pertencentes ao mesmo grupo.

grupo	item	pequena descrição do item
1	5	velocidade positiva em um instante
	18	velocidade negativa em um instante
	13	velocidade mais negativa em um intervalo
2	7	aceleração positiva em um instante
	6	aceleração negativa em um instante
	2	aceleração mais negativa em um intervalo
3	19	procedimento para calcular deslocamento em um intervalo
	4	deslocamento em um intervalo
	23	maior deslocamento em um intervalo
4	10	procedimento para calcular variação de velocidade em um intervalo
	16	variação de velocidade em um intervalo
	1	maior variação de velocidade em um intervalo
5	11	gráfico de velocidade oriundo de um gráfico de posição
	14	gráfico de aceleração oriundo de um gráfico de velocidade
	21	gráfico de posição oriundo de um gráfico de velocidade
	15	gráfico de velocidade oriundo de um gráfico de aceleração
6	8	gráfico de posição: interpretação do movimento particular
	3	gráfico de posição: interpretação do movimento sob velocidade constante
	24	gráfico de velocidade: interpretação do movimento com aceleração constante
	17	gráfico de velocidade: interpretação do movimento com posição crescendo uniformemente
	25	gráfico de aceleração: interpretação do movimento com velocidade crescendo uniformemente
7	9	aceleração positiva e constante: identificar gráfico de posição
	12	velocidade constante: identificar gráficos de posição e velocidade
	22	aceleração constante: identificar gráficos de velocidade e aceleração
	26	velocidade aumentando uniformemente: identificar gráficos de velocidade e aceleração
	20	aceleração crescendo uniformemente: identificar gráfico de aceleração

Tabela 1: Os itens estão separados por grupos, cada um com seu objetivo. Todos os 26 itens apresentam 5 opções de respostas, da letra A a E, sendo apenas uma correta (em negrito) e quatro distratores. A coluna N (nulo) corresponde o percentual de alunos que optaram por não assinalar nenhuma das cinco opções em determinado item.

grupo	item	pré - teste						pós - teste					
		A	**B**	**C**	**D**	**E**	**N**	**A**	**B**	**C**	**D**	**E**	**N**
	5	4	2	**21**	34	38	2	7	3	**20**	62	8	0
1	**18**	**13**	16	11	13	43	5	**20**	25	11	25	13	6
	13	**37**	39	2	17	5	0	**52**	24	4	16	3	2
	7	**15**	9	17	38	17	6	**19**	15	43	11	8	3
2	**6**	20	**13**	9	8	44	6	42	**19**	12	17	9	1
	2	1	23	43	2	**28**	2	1	20	35	1	**43**	0
3	**19**	19	**28**	20	23	6	6	3	**54**	14	22	7	0
	4	6	11	44	**8**	28	4	5	13	10	**42**	29	1
	23	1	**13**	24	29	31	2	3	**20**	18	28	29	2
4	**10**	**26**	24	20	13	11	6	**57**	14	5	3	21	1
	16	17	12	46	**13**	10	2	9	9	19	**29**	32	1
	1	**1**	20	3	17	55	3	**7**	16	2	11	63	2
	11	43	23	16	**15**	0	3	25	25	9	**31**	8	1
5	**14**	43	**24**	13	11	7	2	31	**31**	13	15	9	2
	21	9	**16**	35	9	28	2	11	**26**	19	10	31	3
	15	**15**	39	20	9	13	4	**28**	17	25	6	21	3
6	**8**	17	16	23	**28**	17	0	7	17	17	**48**	11	0
	3	26	0	24	**19**	31	0	21	0	38	**29**	12	0
	24	**6**	72	13	6	0	2	**12**	70	14	3	0	1
	17	**15**	17	2	63	2	2	**43**	13	3	39	2	1
	25	3	8	**13**	62	11	2	4	5	**38**	46	6	1
7	**9**	9	69	4	8	**9**	2	12	56	8	3	**21**	1
	12	6	**11**	29	38	16	0	3	**31**	26	25	16	0
	22	15	24	**12**	10	37	2	18	21	**26**	5	28	2
	26	6	**11**	5	55	21	2	5	**31**	4	49	10	1
	20	4	13	**23**	51	6	2	4	23	**20**	38	12	3

Tabela 2: Percentual de acertos em cada grupo no pré-teste e no pós-teste.

grupo	pré - test	pós - test
1	24	31
2	21	27
3	16	39
4	13	31
5	18	39
6	16	34
7	13	26

distribuição das marcações dos alunos, em termos percentuais, nas duas etapas. Assim, é até mesmo possível verificar visualmente a eficiência dos distratores nos itens e comparar os resultados do pré-teste com o pós-teste.

A seguir, a Tabela 2 expõe de maneira macro, o desempenho dos estudantes nas duas etapas, mostrando o percentual de acertos em cada grupo.

A partir desses valores é possível concluir que os estudantes encontraram grande dificuldade durante a primeira etapa. Já no pós-teste, após o assunto ter sido abordado nas aulas, a média de acertos aumentou significativamente em todos os grupos. Somente nos grupos 1 e 2 o percentual pós-teste não foi pelo menos o dobro do resultado obtido no pré-teste.

Uma vez mapeadas as marcações de cada um dos estudantes nos dois momentos, item por item, deu-se início a estatística descritiva dos resultados do teste. Visando descobrir a eficiência dos distratores, foram calculadas as médias globais no pré e no pós-teste, os desvios padrão e os valores da estatística z, que retrata a diferença entre duas médias amostrais de distribuições normais em unidades de desvio padrão da média (ou erro padrão da média). Sendo assim, para este estudo, o valor z informa a diferença entre a média amostral dos estudantes nos testes e a média de acertos casuais. Os resultados desses parâmetros estatísticos estão expostos nas Tabelas 3 e 4.

Uma vez calculado o valor z, é possível obter, com auxílio da tabela padronizada de áreas associada à distribuição normal, o nível de significância dado pelo valor-p [7]. Quanto menor o valor-p, maior a confiança de se rejeitar a hipótese nula, isto é, rejeitar que as médias em ambos os testes são iguais a média correspondente à chance de acerto casual, caso as alternativas fossem igualmente prováveis, ou seja, 1/5 de 26 itens, portanto um escore total de 5,2 acertos. Em geral, na maioria das análises estatísticas, assim com as realizadas neste estudo, para p <0,05, a hipótese nula pode ser rejeitada.

Conforme apresentado na Tabela 3, o valor de z na 1ªaplicação do teste foi -2,62. A partir dessa informação, pode-se rejeitar a hipótese nula com p <0,01. Observa-se ainda que a média obtida (4,31) foi menor que média atribuída à chance de acerto casual (5,2).

Esse resultado evidencia que os itens não foram respondidos ao acaso. De fato, os estudantes responderam o teste tentando resolvê-lo e por isso foram atraídos por alguns distratores. É possível chegar à essa conclusão não somente pela média do total de acertos abaixo da média que configura o acerto por mero acaso, mas tam-

Tabela 3: Média (M), Variância (V) e Desvio Padrão (DP) de cada item na primeira etapa (pré-teste). Ao lado, estão apresentados esses parâmetros analisando o teste por inteiro, a Média Global (MG), Desvio Padrão da Média ou Erro Padrão da Média (DPM) e o valor z.

	1	2	3	4	5	6	7	8	9	10	11	12	13		
M	0,01	0,28	0,19	0,08	0,21	0,13	0,15	0,28	0,09	0,26	0,14	0,12	0,37		
V	0,01	0,20	0,15	0,07	0,17	0,11	0,13	0,20	0,09	0,19	0,12	0,10	0,23		
DP	0,09	0,45	0,39	0,27	0,41	0,33	0,36	0,45	0,29	0,44	0,35	0,32	0,48	**MG**	**4,31**
	14	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**DP**	**0,33**
M	0,24	0,15	0,12	0,15	0,13	0,28	0,23	0,16	0,12	0,13	0,06	0,13	0,11	**Z**	**-2,62**
V	0,19	0,13	0,10	0,13	0,11	0,20	0,18	0,13	0,10	0,12	0,06	0,12	0,10		
DP	0,43	0,36	0,32	0,36	0,33	0,45	0,42	0,37	0,32	0,34	0,24	0,34	0,31		

Tabela 4: Média (M), Variância (V) e Desvio Padrão (DP) de cada item na segunda etapa (pós-teste). Ao lado, estão apresentados esses parâmetros analisando o teste por inteiro, a Média Global (MG), Desvio Padrão da Média ou Erro Padrão da Média (DPM) e o valor z.

	1	2	3	4	5	6	7	8	9	10	11	12	13		
M	0,07	0,43	0,29	0,42	0,20	0,19	0,19	0,48	0,21	0,57	0,31	0,31	0,52		
V	0,06	0,25	0,20	0,24	0,16	0,15	0,16	0,25	0,17	0,25	0,22	0,21	0,25		
DP	0,24	0,50	0,45	0,49	0,40	0,39	0,40	0,50	0,41	0,50	0,47	0,46	0,50	**MG**	**7,98**
	14	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**DP**	**0,45**
M	0,31	0,28	0,29	0,43	0,20	0,54	0,20	0,26	0,26	0,20	0,12	0,38	0,31	**Z**	**6,17**
V	0,22	0,20	0,21	0,25	0,16	0,25	0,16	0,20	0,21	0,17	0,11	0,24	0,21		
DP	0,47	0,45	0,45	0,50	0,40	0,50	0,40	0,44	0,45	0,41	0,33	0,49	0,46		

 DOI: http://dx.doi.org/10.1590/1806-9126-RBEF-2019-0149

bém observando as distribuições de frequência nas cinco alternativas que, conforme está na Tabela 1, não configuram equiprobabilidade das alternativas. Poucos foram os itens onde a opção correta obteve mais marcações que os distratores.

Na 2ªaplicação (pós-teste) o valor de z saltou para 6,17, sendo assim possível rejeitar a hipótese nula com p <0,001. Nota-se também que a média obtida no nessa etapa (7,98) foi estatisticamente superior à média esperada para o acerto casual (5,2), o que demonstra um pequeno efeito positivo da intervenção do professor, por meio de aulas tradicionais.

Uma vez realizada a estatística descritiva, faz-se necessário realizar o procedimento estatístico da validação do teste. Essa estatística é composta por quatro parâmetros escolhidos dentre os cinco índices empregados por Zavala para validar a nova versão do TUG-K: i) índice de dificuldade de item (P); ii) coeficiente de ponto bisserial (r_{bps}); iii) índice de confiabilidade de Kuder-Richardson (KR – 20); iv) delta de Ferguson (δ). Assim, poderemos comparar os dados obtidos de estudantes de ensino médio no Brasil com calouros universitários no México.

Optamos neste trabalho por suprimir o índice de discriminação do item (D_i) que foi empregado por Zavala em sua validação do TUG-K. Essa escolha foi motivada pelo fato de que o referido índice é uma estatística alternativa ao coeficiente de correlação item total (coeficiente de ponto bisserial), criada em uma época em que não existiam computadores e em que era importante economizar cálculos através de atalhos, expressos por fórmulas mais simples. Ou seja, o índice de discriminação não agrega qualquer informação relevante adicional a dada pelo coeficiente de correlação item-total, além de possuir as seguintes restrições quanto à sua utilidade: 1 – não é adimensional como o coeficiente de correlação item-total; 2 – seus valores mudam conforme se modifica a extensão dos grupos extremos. 3 – a relação dele com o coeficiente de correlação item-total depende da forma da distribuição do escore total [8-10].

No Quadro 3 encontram-se as breves descrições acerca dos significados dos parâmetros estatísticos usados na análise dos resultados, assim como seus valores desejados.

O que verdadeiramente fundamenta a validação da versão atualizada do TUG-K, traduzida para o português, no ensino médio, é justamente a interpretação dos valores calculados desses parâmetros supracitados, tendo como referência os valores esperados (Quadro 3) e os valores medidos por Zavala.

As Tabelas 5 e 6 expõem os valores encontrados dos três primeiros parâmetros expostos no Quadro 3, calculados item por item, após a aplicação do pré e do pós-teste. Já a Tabela 7 apresenta os resultados desses mesmos parâmetros obtidos por Zavala.

Com base nessas tabelas, foram gerados os gráficos das Figuras 1 e 2, que permitem um comparativo mais claro entre os resultados.

Pode-se perceber, após uma leitura dos gráficos dos índices de dificuldade apresentados na Figura 1, que os

Quadro 3: Descrição dos quatro parâmetros estatísticos calculados após a aplicação do TUG-K. Os dois primeiros são calculados item por item, enquanto os dois últimos são parâmetros globais, envolvendo o teste inteiro.

Parâmetro	Significado	Valores limites	Valores desejados
Índice de dificuldade (P)	Percentual de acertos do item e, portanto, deve ser interpretado de maneira complementar ao seu valor, i.e. quanto maior, o percentual de acertos do item, maior será o valor de P e, consequentemente, menor será a dificuldade.	[0, 1]	[0.3, 0.9]
Coeficiente de ponto bisserial (r_{bps})	Fiabilidade ou fidedignidade de um único item do teste, definido como o coeficiente de correlação entre os escores no item e as respectivas pontuações totais do teste, para todos os alunos. Um item com boa fiabilidade é um item que os alunos com maiores escores tenham um desempenho superior aos alunos com escores baixos. Um item deve ser retirado do teste se tiver esse coeficiente negativo.	[-1, 1]	$\geq$ 0.2
Índice de confiabilidade Kuder – Richardson (KR – 20)	Trata-se da consistência interna total do teste. Quanto mais elevadas forem as correlações entre os itens, maior é a homogeneidade dos itens e maior é a consistência com que medem o constructo teórico. A consistência interna é o parâmetro mais relevante para assegurar a fiabilidade ou fidedignidade global do teste, isto é, se um instrumento de medida fornece sempre os mesmos resultados quando aplicado a alvos estruturalmente iguais.	[0, 1]	$\geq$ 0.7
Delta de Ferguson (δ)	É a capacidade de discriminação do teste como um todo através da estimativa da amplitude da dispersão dos escores totais de todos os participantes.	[0, 1]	$\geq$ 0.9

Tabela 5: Índice de dificuldade (P) e o coeficiente de ponto bisserial (r_{pbs}) de cada item na primeira aplicação do TUG-K no ensino médio.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
P	0,01	0,28	0,19	0,08	0,21	0,13	0,15	0,28	0,09	0,26	0,15	0,11	0,37
rpbs	-0,03	0,56	0,46	0,12	0,50	0,49	0,47	0,60	0,43	0,23	0,51	0,55	0,58
	14	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**
P	0,24	0,15	0,13	0,15	0,13	0,28	0,23	0,16	0,12	0,13	0,06	0,13	0,11
rpbs	0,49	0,29	0,22	0,65	0,36	0,29	0,07	0,26	0,54	0,32	0,30	0,51	0,47

Tabela 6: Índice de dificuldade (P) e o coeficiente de ponto bisserial (r_{pbs}) de cada item na segunda aplicação do TUG-K no ensino médio.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
P	0,07	0,43	0,29	0,42	0,20	0,19	0,19	0,48	0,21	0,57	0,31	0,31	0,52
rpbs	0,45	0,39	0,58	0,44	0,47	0,43	0,44	0,59	0,51	0,22	0,48	0,50	0,52
	14	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**
P	0,31	0,28	0,29	0,43	0,20	0,54	0,20	0,26	0,26	0,20	0,12	0,38	0,31
rpbs	0,39	0,34	0,30	0,53	0,25	0,41	0,11	0,25	0,48	0,48	0,48	0,56	0,53

Tabela 7: Índice de dificuldade (P) e o coeficiente de ponto bisserial (r_{pbs}) de cada item obtido por Zavala, após a aplicação do TUG-K em 471 estudantes de um curso introdutório de física em uma universidade no México.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
P	0,11	0,62	0,58	0,39	0,39	0,46	0,40	0,51	0,46	0,55	0,45	0,48	0,51
rpbs	0,47	0,42	0,51	0,48	0,43	0,46	0,42	0,58	0,57	0,40	0,51	0,63	0,60
	14	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**
P	0,54	0,46	0,33	0,70	0,41	0,67	0,52	0,50	0,48	0,33	0,32	0,54	0,51
rpbs	0,55	0,50	0,42	0,59	0,43	0,45	0,32	0,44	0,58	0,43	0,60	0,61	0,65

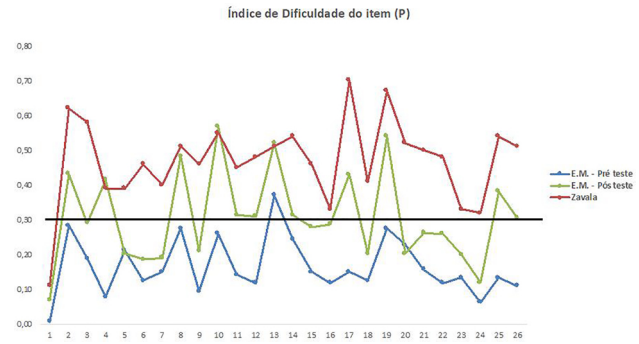

Figura 1: O gráfico em azul representa o índice de dificuldade de cada item para os 127 estudantes de ensino médio na primeira aplicação (pré-teste). O gráfico verde, por sua vez, mostra o mesmo parâmetro, na segunda aplicação do teste. Já o gráfico em vermelho é o resultado publicado por Zavala. A reta horizontal fixada em P = 0,30 serve com uma linha de referência. Assim, os valores acima dessa linha são considerados desejáveis.

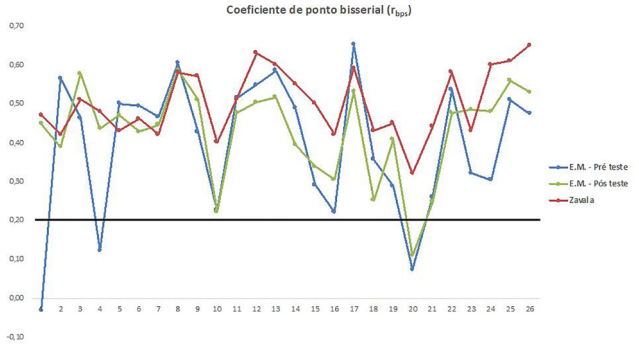

Figura 2: Mantendo o padrão de cores dos dois gráficos anteriores, em azul e em verde tem-se o coeficiente de ponto bisserial obtido a partir das marcações dos estudantes na 1ª e na 2ª aplicação do teste, respectivamente. E, em vermelho, os resultados de Zavala. A linha de referência está fixada em $r_{bps} = 0{,}20$.

alunos acertaram poucos itens no pré-teste, corroborando a conclusão feita previamente com o auxílio das Tabelas 4 e 5. Em relação ao coeficiente de ponto bisserial, a Figura 2 mostra visualmente um comportamento bem similar entre os valores de Zavala e os resultados obtidos com os alunos brasileiros.

Fica evidente da Figura 1 que apenas o item 01, cujo enunciado está na Figura 3, se mostrou muito mais difícil que o desejado, levando-se em consideração tanto os estudantes de ensino médio brasileiros quanto universitários em um curso introdutório de física no México.

Em nossa pesquisa, o item 01 apresentou apenas 1 marcação correta entre os 127 estudantes no pré-teste. O desempenho nesse item no pós-teste também não foi muito diferente, com somente 7 marcações corretas, sendo 5 dessas pertencentes aos 5 melhores escores. De fato, retornando à Tabela 1, temos que nas duas aplicações do teste, o distrator mais marcado foi a letra E. Uma explicação plausível para o alto percentual de erro nesse item, tanto na aplicação de Zavala quanto na presente pesquisa é terceira dificuldade típica identificada por Beichner em seu artigo, denominada confusão de variáveis [2]. De fato, os estudantes costumam confundir os conceitos de posição, velocidade e aceleração quando representados através gráficos, sem atentar que o "desenho" do gráfico, ou melhor o formato da função descrita no gráfico, depende da grandeza representada no eixo vertical. É muito comum, por exemplo, o estudante atribuir um gráfico posição por tempo contendo uma reta inclinada crescente a um objeto cuja velocidade aumenta, devido à confusão de variáveis. Nesse item específico, essa dificuldade de confusão de variáveis pode ter sido potencializada por conta do termo "maior variação da velocidade", que aparece explicitamente no enunciado, isto é, os estudantes escolheram o gráfico da letra E que mostra uma função que "varia mais", confundindo o eixo vertical que representa a aceleração com a velocidade. Uma sugestão para melhorar o TUG-K seria trocar esse item de lugar, deixando de ser logo o primeiro item do teste, em virtude de sua complexidade ao exigir o cálculo da área em

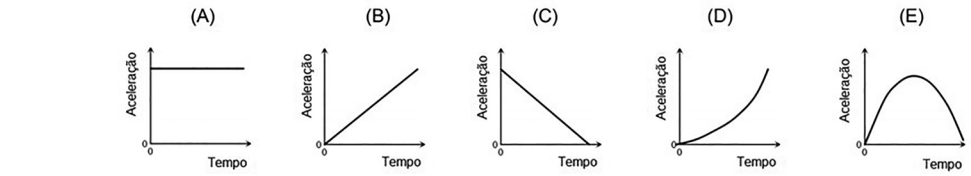

Figura 3: O primeiro item do teste e, justamente aquele de menor percentual de acerto tanto na presente pesquisa quanto no artigo do Zavala.

um gráfico de aceleração por tempo. Possivelmente com um efeito modesto sobre a dificuldade do item, podemos ainda sugerir a alteração do termo "maior variação de velocidade" por "maior mudança no valor da velocidade".

Em relação ao segundo parâmetro analisado, o coeficiente bisserial (r_{bps}), 63 estudantes (praticamente a metade dos 127) obtiveram desempenho superior ao do aluno que acertou o primeiro item e, por esse motivo, o coeficiente de ponto bisserial desse item apresentou valor negativo. Os itens 4 e 20 também apresentaram baixíssima fiabilidade no pré-teste, mas o item 4 teve bom resultado nesse quesito no pós-teste.

No caso do item 20 (Figura 4), esse índice foi muito baixo em ambos os testes, apresentando um percentual de acertos até ligeiramente menor na segunda aplicação do teste. Além disso, esse foi o único item que apresentou um coeficiente de ponto bisserial abaixo de ideal, mesmo após as aulas. No estudo de Zavala, esse item também teve o menor valor do coeficiente bisserial comparado aos outros, no entanto o seu valor ainda manteve-se acima do valor desejável para esse tipo de coeficiente, o que mostra a sua adequação do ponto de vista estatístico nessa aplicação.

O distrator mais marcado desse item, em ambas as etapas, foi a letra D, provavelmente por ser a opção cujos gráficos são retas que aumentam uniformemente, independentemente de qual grandeza cinemática está no eixo vertical. Novamente, esse resultado possivelmente está associado com a terceira dificuldade catalogada por Beichner, a saber a confusão de variáveis [2]. Realmente, o distrator mais marcado (letra D) apela para a forma funcional dos três gráficos que representa uma reta crescente, isto é algo que aumenta uniformemente conforme o enunciado, mas implicitamente desconsidera que as grandezas representadas no eixo vertical em cada gráfico são distintas. A grande e ampla eficiência desse distrator nesse contexto particular, possivelmente contribuiu para o baixo valor do coeficiente bisserial do item, demonstrando quão resilientes e ubíquas são as concepções alternativas dos estudantes submetidos às aulas tradicionais.

O próximo passo, uma vez analisados os itens do teste, é interpretar os valores extraídos do teste por inteiro, que estão expostos na Tabela 8, a fim de verificar sua validade quando aplicado a estudantes de ensino médio.

Confrontando os quatro parâmetros da Tabela 8 com o Quadro 3, nota-se que a aplicação do TUG-K com

Tabela 8: Resultados da estatística completa do teste, em ambas as etapas. Na última coluna, estão os valores medidos por Zavala.

Parâmetro	CPII	Zavala
Média dos índices de dificuldade	0,17	0,47
Média dos coeficientes bisseriais	0,39	0,50
KR - 20	0,79	0,88
Delta de Ferguson	0,93	0,99

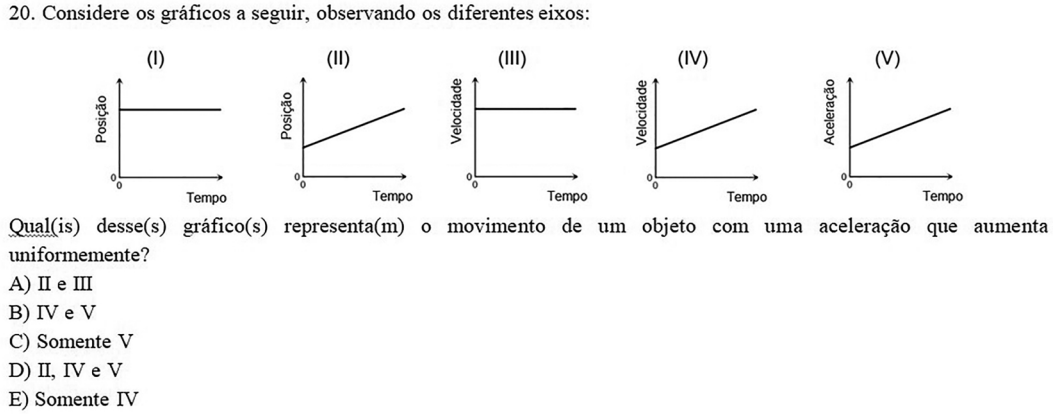

Figura 4: Item 20. O enunciado desse item pede quais gráficos representam um movimento cuja aceleração aumenta uniformemente. Na segunda aplicação, esse item apresentou o menor coeficiente bisserial de todo o teste.

estudantes brasileiros de ensino médio revelou resultados bastante satisfatórios, pois todos os parâmetros estão dentro das faixas de seus respectivos valores desejáveis. A única ressalva é a média dos índices de dificuldade baixa no pré-teste, mas que depois se tornou acima do desejável na etapa pós-teste. O que significa dizer que o teste foi muito difícil para os alunos de ensino médio que não haviam tido contato com o assunto ainda e, mesmo após as aulas, continuou difícil, porém dentro de um valor minimamente aceitável. É de se esperar que universitários apresentem menos dificuldade do que estudantes de ensino médio.

Por sua vez, o valor calculado do Delta de Ferguson também está dentro do desejável, refletindo o fato de que o teste inteiro apresentou boa capacidade de discriminação.

Os itens, de modo geral, apresentaram boa fiabilidade medida pelo coeficiente de correlação item-total (coeficiente bisserial). Esse parâmetro é importante pelo seu impacto na fidedignidade total do escore total do teste que pode ser estimada pela equação KR-20 [11] que também foi utilizada por Zavala e Beichner na versão original do teste [2]. Um coeficiente de fidedignidade elevado corrobora a presunção de existência de um traço latente ou construto comum a todos os itens do teste, ou seja, neste caso, é uma das evidências de que o teste quantifica de forma consistente a compreensão de gráficos da cinemática. O valor obtido para esse índice mostra que esse teste pode ser aplicado ano após ano para diferentes estudantes de ensino médio.

Adicionalmente, foi calculado o ganho de aprendizagem de Hake [12] desse grupo de alunos, muito embora esse índice não se constitua em um parâmetro estatístico e sim numa estimativa quantitativa da aprendizagem. O ganho de Hake é definido como uma medida aproximada da eficácia de uma intervenção didática na promoção da compreensão conceitual e pode ser calculado através da proporção entre a quantidade que os alunos aprenderam dividida pela quantidade que poderiam ter aprendido. Sendo assim, após uma sequência de aulas sobre cinemática centradas no professor, trabalhando o conteúdo de maneira expositiva, o ganho de Hake global, para todos os alunos, foi de aproximadamente 17%. Esse valor é compatível com os resultados disponíveis da literatura na área relativos às aulas tradicionais, no exterior e no Brasil, que costumam apresentar um ganho de Hake entre 10% a 30% [12, 13].

2. Conclusões

Toda a discussão feita anteriormente, desde análise individual dos itens, até a interpretação dos valores obtidos para os parâmetros expostos na Tabela 8, é suficiente para declarar que a versão do TUG-K 4.0, traduzida para o português, demonstrou possuir validade estatística para ser aplicada no ensino médio.

Mesmo com um resultado positivo para a validação estatística do TUG-K com estudantes de ensino médio da rede federal, sugerimos alguns pontos que precisam ser objeto de investigação mais profunda com o intuito de incrementar os valores dos parâmetros estatísticos relacionados com a confiabilidade do teste. Em especial, indicamos que os itens 01 e 20 passem por uma reavaliação de validade de conteúdo por especialistas e sejam submetidos novamente a uma validação estatística com um espaço amostral similar ou maior do que o presente trabalho, com estudantes de ensino médio. A princípio, a mera exclusão desses itens melhora o parâmetro de fiabilidade global do teste, porém eles devem ser substituídos por itens equivalentes para não comprometer a distribuição e a simetria de questões entre os 7 grupos de objetivos do TUG-K. Além disso, caso se opte pela manutenção do item 01, com seu elevado grau de dificuldade, então sugerimos que ele seja deslocado para o meio do teste, momento em que os estudantes já estão mais concentrados e já resolveram questões semelhantes e mais complexas que envolvem o cálculo da área de um gráfico.

A perspectiva subjacente a este trabalho é difundir e incentivar o uso, junto aos professores de ensino médio, de um instrumento eficiente e confiável de avaliação tanto das concepções alternativas dos estudantes sobre gráficos de cinemática, quanto dos seus níveis de aprendizado conceitual e operacional do tema.

Referências

[1] `https://www.physport.org/assessments/assessment.cfm?I=6&A=TUGK`, acessado em 15/06/2019.

[2] R.J. Beichner, American Journal of Physics **62**, 750 (1994).

[3] L.C. Mcdermott, M.L. Rosenquist e E.H. van Zee, American Journal of Physics **55**, 503 (1986).

[4] J.R. Mokros e R.F. Tinker, Journal of Research in Science Teaching **24**, 369 (1987).

[5] G. Zavala, S. Tejeda, P. Barnio e R.J. Beichner, Physical Review Physics Education Research **13**, 020111 (2017).

[6] D.A. Agrello e R. Garg, Revista Brasileira de Ensino de Física**21**, 103 (1999).

[7] W.O. Bussab e P.A. Morettin, in: *Estatiìstica Baìsica* (Atual Editora, São Paulo, 1988).

[8] F.L. Silveira, Ciência e Cultura **33**, 146 (1981).

[9] R.J. Wherry, in: *Contributions to correlational analysis* (Academic Press, Orlando, 1984).

[10] F.L. Silveira, Ciência e Cultura **32**, 214 (1980).

[11] G.F. Kuder e M.W. Richardson, Psychometrika **2,** 151 (1937).

[12] R.R. Hake, Am. J. Phys **66**, 64 (1998).

[13] M.P. Quibao, A.C. Silva, N.S. Almeida, R.M.A.A. Silva, S.R. Muniz e F.F. Paiva, Rev. Bras. Ensino Fís., **41**, e20180258 (2019).