Test of understanding graphs in kinematics: Item objectives confirmed by clustering eye movement transitions

P. Klein,[1,*] S. Becker,[2] S. Küchemann,[2] and J. Kuhn[2]

[1]Faculty of Physics, Physics Education Research, Georg-August-Universität Göttingen, Friedrich-Hund-Platz 1, 37077 Göttingen, Germany

[2]Department of Physics, Physics Education Research Group, Technische Universität Kaiserslautern, Erwin-Schrödinger-Straße 46, 67663 Kaiserslautern, Germany

(Received 7 December 2020; accepted 9 February 2021; published 17 March 2021)

The test of understanding graphs in kinematics (TUG-K) has widely been used to assess students' understanding of this subject. The TUG-K poses different objectives to the test takers such as (1) the selection of a graph from a textual description, (2) the selection of corresponding graphs, and (3) the selection of a textual description from a graph. Whether test takers follow these task requirements is usually inferred from evaluating the test scores as correct or incorrect, yet the process of how students actually interact with the different tasks remains unknown. Recent studies have shown that eye tracking can provide rich insight into student's interaction with multiple-choice tasks. In the current work, we analyzed the eye movement patterns of $N = 115$ high school students while solving the TUG-K. Each question was divided into a question area (Q) and an option area (O), then gaze transitions between Q and O and between different options were calculated. A cluster analysis using the transition metrics revealed three item groups, containing the aforementioned objectives of the items. The clusters remain stable for different subsamples of our dataset, for instance, considering only the correct or only the incorrect responses, or considering high- or low-confidence responses. We conclude that eye movements can reflect task demands on a procedural level well beyond the classical methods of evaluating test scores, eventually making eye tracking an additional method for item analysis that can be utilized to confirm or explore test and item structures.

DOI: 10.1103/PhysRevPhysEducRes.17.013102

I. INTRODUCTION

Since the development of the Force Concept Inventory (FCI) [1], the design and use of research-based distractor-driven multiple-choice items has accelerated. The PhysPort collection comprises more than 90 research-based assessments for introductory and upper-level physics concepts, scientific reasoning, problem solving, or student attitudes and beliefs [2]. The methodologies utilized to create and evaluate these assessments include student interviews, expert reviews, and statistical analysis of the instruments' psychometric properties. However, the way in which the test participants visually interact with the tasks has hardly been taken into account so far, especially not for the purpose of a systematic instrument or item analysis. Previous studies—one using the FCI, the others using the test of understanding graphs in kinematics (TUG-K) [3]—have examined the test takers' visual attention distribution on the stems and options of the test items [4–7]. In both studies, the domain experts not only achieved better test scores, they were also more efficient and effective in distributing their visual attention to the distractors by avoiding naive or intuitive ideas (i.e., alternative conceptions) and focusing on the correct alternatives. Both studies showed that using a subject's eye movements can provide insight into elements of their performance beyond simple totals of correct and incorrect responses. Furthermore, the analysis of eye movements can reveal whether students approach test items in problem-solving mode rather than merely recalling declarative information [8,9].

Here, we demonstrate that capturing students' visual attention while solving (physics) concept tests also has the potential to confirm test and item structures at the level of visual behavior. Particularly for the different item formats of the TUG-K, visual attention can provide information on whether the items are processed as intended by the test makers, for example, whether test takers actually compare the different graphs in the options.

II. STATE OF RESEARCH

The TUG-K, introduced by Beichner in 1994 [3], has become one of the most widely used tests to date designed

*pascal.klein@uni-goettingen.de

TABLE I. Overview of TUG-K item objectives and representational formats.

Item format (Question, Options)	Items	Objectives
Text → Graph	1, 23	(1) Identify a graph with greatest change in a variable
	9, 12, 20, 22, 26	(2) Select a graph from a textual description
Graph → Graph	11, 14, 15, 21	(3) Select corresponding graph from a graph
Graph → Text	3, 8, 17, 24, 25	(4) Select a textual description from a graph
	10, 19	(5) Establish the procedure to determine the change of a variable
Graph → Value	2, 5, 6, 7, 13, 18	(6) Evaluate the slope of a graph
	4, 16	(7) Evaluate the area under the curve

to evaluate students' understanding in kinematics. It addresses several objectives that are summarized in Table I. In its recently modified version 4.0 [10] the test comprises 26 items that were created based on the extensive research on student difficulties with graphs of position, velocity, and acceleration versus time. Besides the mathematical evaluation of kinematic quantities from a graph (objectives 6 and 7 in Table I), the items pose various demands on the test takers, especially to organize and integrate the information that is presented. Some items start from a textual description and require the student to select a graph (objective 2), or vice versa (objective 4), while other items require a mapping between two corresponding graphs (objective 3). The task objectives, for which such selection processes are central, are especially accessible to an eye-tracking analysis. The students must integrate different information from the questions and options, and the eye movements can reflect these integration processes. All TUG-K items will be included in our analysis; however, we are particularly interested in the attentional patterns when the test takers are facing the demands postulated in objectives 2–4. As we show in the next section, some eye-tracking studies already exist for multiple-choice tests in physics education research (PER); however, a grouping of test items based on apparently similar requirements has not yet been done.

III. CAPTURING VISUAL ATTENTION WITH EYE TRACKING

In various PER studies, visual attention was investigated to gain process-related insights into problem solving. Most recently, Susac *et al.* investigated the students' attentional distribution on optical interference and refraction patterns that were presented as response options in multiple-choice tasks [11]. They evaluated dwell times on the different options and correlated them with response accuracy. The test takers attention on the different answer options was also important in the study by Han *et al.* who found that students pay a lot of attention to wrong options despite correct answers [4]. They concluded that alternative force conceptions are persistent [4]. In a study on the TUG-K, we also found that although attention to popular (wrong) options decreases with increasing expertise, it is still present [5]. The early studies from Madsen *et al.* [12], the recent studies from Viiri *et al.* [13], and the studies from Klein *et al.* [14] should also be mentioned in this context. In these studies, students' attention on different response options were used, again evaluating fixation counts or visit durations (viewing times). When multiple-choice tasks are being processed, the attention naturally changes between the item stem and response options, and oftentimes the different response options are considered. Therefore, it is surprising that transitions have not been considered in the studies mentioned. A recent investigation by Viiri *et al.* examined the number of transitions from the stem to the options when solving multiple-choice problems including different representations (see, e.g., text and graphs) [15]. With a rather small sample size ($N = 8$), the study indicated that the transitions between the text (stem) and the graph (options) were different for students who prefer either the text or graph representation. They interpreted the transitions between the question and the options as the number of times a student must reread the stem for connecting it to the options and how difficult (or easy) it is to remember the stem. The transitions between the different options were related to decision-making processes and comparisons between different options. Especially when different forms of representations such as text and diagrams have to be integrated with each other, as is the case with the TUG-K, transitions can provide information about the students' coordination behavior [16]. Because the item objectives listed in Table I involve highly visual processes (e.g., selecting and integrating information), the goal of our study is to further exploit eye tracking as a method for item analysis on a behavioral level, particularly using transition metrics that reflect an important aspect of the test takers response process. By gathering eye gaze data from a exceptionally large sample (compared to other eye-tracking studies [17]), we aim to answer the following research questions. (1) Do eye gaze transitions based on item stems and options reveal task objectives for TUG-K items? (2) If so, do we find stability of the result regarding different subsamples, for example, considering only correct answers?

IV. METHODS

The data were collected using stationary eye-tracking systems installed in the schools' libraries.

Participants and data collection.—In total, 115 high school students (58 female and 57 male, all with normal or correct-to-normal vision, all German native) took part in the study. At the time of the test, the subject kinematics was completed in all classes.

Material.—The items of the latest version of the TUG-K were presented to the students in the original order of the test in German. Besides gathering students' responses, their response confidence was assessed using a four-point rating scale for each item.

As reported by the test constructors, the Kuder-Richardson reliability index exceeds the desired benchmark of 0.7 for group measures ($\alpha = 0.88$ for the modified version) and the average point-biserial coefficient is $r = 0.50$ [10]. For our dataset, we obtained $\alpha = 0.90$ and $r = 0.47$, and the students' mean test score and standard deviation were ($59 \pm 25\%$), ranging from 4% (1 question correct, 1 student) to 100% (26 questions correct, 2 students).

Apparatus.—The eye movements were recorded using a Tobii Pro X3-120 stationary eye-tracking system, operating with an accuracy of less than 0.40° of visual angle and a sampling frequency of 120 Hz. An eye movement was classified as a saccade (i.e., in motion) if the acceleration of the eyes exceeded $8500°/s^2$ and velocity exceeded $30°/s$. The average distance between the eyes of the students and the screen was 62 cm.

Data preparation.—For each item, two areas of interest (AOIs) were defined; see Fig. 1. The Q-AOI covers the question (item stem) which consists of text or a graph. The O-AOI covers all (five) answer options, consisting of written statements, values, or graphs. The O-AOI was divided into five smaller AOIs, each covering the single choices, for calculating the transitions between different answer options. Figure 1 also illustrates the eye movement data from one participant.

Cluster analysis.—To identify item groups, a hierarchical cluster analysis using agglomerative methods based on the squared Euclidean distance was performed [18], taking into account two measures of eye movements: the transitions between the question and options, and the transitions between different options. To select the method of clustering, the agglomerative coefficient was calculated, indicating the strength of the clustering structure. To select the number of clusters, the average silhouette approach was used [19]. The cluster analysis was conducted via the R-package cluster (version 2.1.0) [20], visualized with FACTOEXTRA [21], and two dendrograms were compared using DENDEXTEND [22]. For the latter, the quality of the alignment of two dendrograms can be quantified by the entanglement measure, ranging from 0 (no entanglement) to 1 (full entanglement). A lower entanglement coefficient corresponds to a good alignment [22].

V. RESULTS

The Ward method was assessed to provide the strongest cluster structure, yielding an agglomerative coefficient of 0.95, and outperforming other agglomerative methods (average linkage 0.87, single linkage 0.80, complete linkage 0.91). The results of hierarchical clustering are presented in a dendrogram; see Fig. 2. The bottom column (26 nodes) represents the initial data (items), and the remaining nodes represent the clusters to which the items belong, with the vertical lines representing the distance (dissimilarity). The height of each node in the plot is proportional to the value of the intergroup dissimilarity between its two daughters. Inspecting the dendrogram, we find a three-cluster solution that was also confirmed using the average silhouette approach. The first cluster (referred to as cluster A hereafter) includes the items 1, 9, 12, 20, 22, 23, and 26, the second cluster (cluster B) includes the items 11, 14, 15, and 21, and the third cluster (cluster C) includes all remaining items. The solution is also presented in Fig. 3, illustrating how the clusters are characterized by the transition metrics. The cluster A is characterized by items

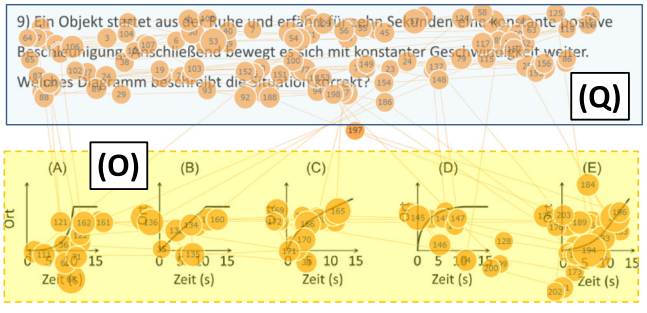

FIG. 1. Definition of AOIs. The Q-AOI covers the question and the O-AOI covers all answer choices. The AOIs that cover single options are not shown for the sake of clarity.

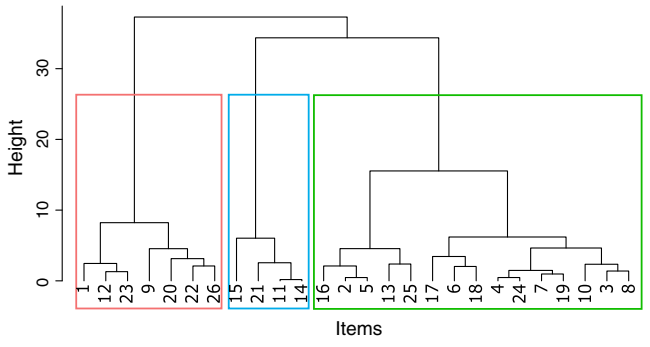

FIG. 2. Dendrogram illustrating the arrangement of the clusters produced by the hierarchical cluster analysis using the Ward method.

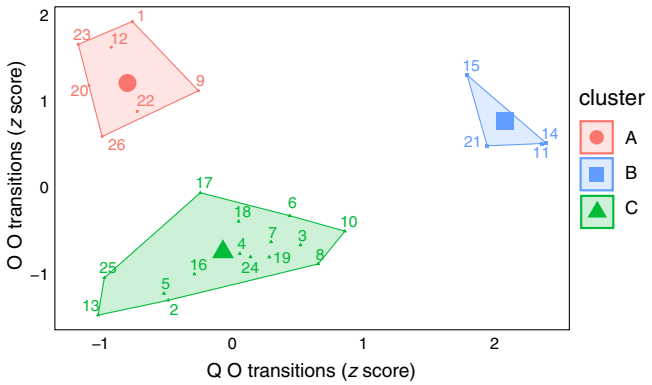

FIG. 3. Characterization of the three clusters emerging from cluster analyses based on eye movement transition measures during problem solving.

with a low number of transitions between question and options (QO transitions) but high number of transitions between options (OO transition). The items in cluster B have both high QO and OO transitions, and the cluster C is defined by items with low numbers of QO and OO transitions. In a next step, we used inference statistical methods to check whether the interaction of OO and QO transitions exhibits a statistically significant difference between the clusters. A 2×3 analysis of variance (ANOVA) with type of transitions (QO, OO) as a within-item factor and cluster (A, B, C) as a between-item factor was performed. We found a significant main effect of the factor cluster [$F(2, 23) = 31.5$, $p < 0.001$, $\eta_p^2 = 0.73$] and a significant interaction between cluster and transition type, $F(2, 23) = 50.6$, $p < 0.001$, $\eta_p^2 = 0.82$.

Note that all data were included in the analysis, regardless of whether the students responded correctly or incorrectly. In a next step, the transition metrics were additionally calculated using only correct and incorrect responses to the questions, and the same procedure as described above was applied. In the Supplemental Material [23], two dendrograms are presented side by side, with the labels connected by lines. The entanglement between both solutions is very low (0.08), meaning that the dendrograms are very similar. For the incorrect responses, also a three-cluster solution was obtained (entanglement 0.08) with only one item (21) being assigned to two different clusters. As above, the 2×3 ANOVAs for correct and incorrect responses yield strong interaction effects between cluster and transition type; see Table II. That is, the three clusters can be distinguished significantly using the QO and OO transition metrics.

Last, we also used the confidence responses to restrict our dataset and to recalculate the transition means per item. Because the students reported their response confidence on a rating scale, we first performed an item-based median split into low- and high-confidence responses, aggregated the transition metrics for the resulting datasets accordingly, and, finally, used cluster analysis as before. The three-cluster solution remained stable, and the ANOVA results are presented in Table II, lining up with the results of the other data splits. So, independent of the split criterion, we find that the number of transitions differs between the clusters, and there is an interaction between transition type and cluster.

VI. DISCUSSION

Cluster A includes the items that belong to the objectives 1 and 2 in Table I, starting from a text and offering graphs as options. When solving these items, students perform many transitions between the different options, and few transitions between the question and the option. If a graph has to be selected from a set of many graphs, then these graphs were compared more often with each other than textual descriptions or values were compared with each other, possibly due to different problem-solving heuristics. Additional verbal data are required to finally support this conclusion. In general, transitions are performed for comparison purposes and for relating given information to each other [15]. For comparing graphs, the pieces of information have to be visually encoded piece by piece, resulting in more transitions based on the limited capacity of the visual working memory [24]. In line with this reasoning, adding graphs to the question should increase the number of transitions between question and the options. And indeed, cluster B is classified by the highest number of transitions between the question and the option, and between different options. Cluster B includes all items that require the

TABLE II. Number of eye movement transitions for subsamples of the dataset and ANOVA results.

	Cluster						2×3 ANOVA statistics								
	A		B		C		Transition type			Cluster			Interaction		
Data included	QO	OO	QO	OO	QO	OO	F	p	η_p^2	F	p	η_p^2	F	p	η_p^2
All responses	6.3	17.5	22.1	14.2	11.0	6.6	0.2	n.s.	⋯	31.5	0.000	0.73	50.6	0.000	0.81
Correct responses	6.1	17.3	20.2	13.7	11.0	6.4	0.0	n.s.	⋯	23.1	0.000	0.67	39.2	0.000	0.77
Incorrect responses	6.5	17.7	26.2	15.7	11.3	7.0	2.5	n.s.	⋯	29.0	0.000	0.72	63.6	0.000	0.85
High-confidence responses	5.2	14.8	14.8	10.0	8.3	5.0	0.6	n.s.	⋯	19.4	0.000	0.63	51.0	0.000	0.82
Low-confidence responses	6.7	18.3	24.1	15.4	12.4	7.4	0.6	n.s.	⋯	29.2	0.000	0.72	47.3	0.000	0.80

n.s. stands for (statistcally) "not significant," indicating p-values $> .05$.

selection of a corresponding graph from a graph (objective 3 in Table I). The graphs in the answer choices are often similar in several sections, so that a comparison of each individual graph in the answer choices with the graph in the question is necessary. Accordingly, this process requires an increased number of visual transitions, or integration processes, between the graph in the question and the graph in the answer choices. The cluster C items show the fewest transitions between the question and the answer choices. All items in this cluster start from a graph and require text or values as outputs. The cluster analysis did not distinguish between different objectives (4)–(7), that means that the transitional behavior is similar between evaluating values and choosing a textual description. When evaluating the graph, for example, calculating slopes or areas, the visual attention is typically focused on the graph to apply a procedure, and after evaluation, the options are considered to match the (mental) evaluation. Note that when restricting our analysis to the items that reflect the selection processes given in objectives 2,3, and 4 (items 3, 8, 9, 11, 12, 14, 15, 17, 20, 21, 22, 24, 25), we find them separated in the three different clusters.

Interestingly, we note from Table II that there is a difference regarding the transition metrics between high- and low-confidence responses, and the difference is much bigger than between the correct and incorrect responses. Concerning cluster A, for instance, the high-confidence responses were related to 5.2 QO transitions and 14.8 OO transitions (on average), whereas the low-confidence responses were related to 6.7 QO and 18.3 OO transitions, thus reflecting differences of +29% and +23.6%, respectively. The gap between correct and incorrect responses is consistently smaller. This result indicates that gaze transitions are more strongly correlated with confidence than with accuracy, a notable result that can be explored in follow-up studies.

VII. CONCLUSION

Here, two transition measures for TUG-K items were determined based on a comparable huge dataset of $N = 115$ high school students. The measures were subjected to a cluster analysis, yielding three statistically separable clusters, A, B, and C, that reflect different combinations of graphs and text or values in the questions and options. The clustering structure has also been proven stable compared to the evaluation of partial datasets, thus allowing an interpretation predominantly independent of the sample. Thus, the eye movements reflected task demands on a procedural level well beyond the classical methods of evaluating test scores, eventually making eye tracking an additional method for supporting (normative) test structures and item analysis. The AOIs are based on superficial test structures (questions and options) making this method suitable to be generalized to other inventories. Deeper insight into students' responses, however, likely requires a detailed look at the actual content that is presented in the questions and options beyond the level of these surface features.

We see the application of cluster analysis to eye-tracking data in general as a promising method to gain insight not only into different types of learners or reading strategies, but also into the procedural requirements of task types.

ACKNOWLEDGMENTS

We thank Andreas Lichtenberger for the help with data collection. We acknowledge support by the Open Access Publication Funds of Göttingen University.

[1] D. Hestenes, M. Wells, and G. Swackhamer, Force Concept Inventory, Phys. Teach. **30**, 141 (1992).
[2] See https://www.physport.org/assessments/.
[3] R. J. Beichner, Testing student interpretation of kinematics graphs, Am. J. Phys. **62**, 750 (1994).
[4] J Han, L. Chen, Z. Fu, J. Fritchman, and L. Bao, Eye-tracking of visual attention in web-based assessment using the Force Concept Inventory, Eur. J. Phys. **38**, 045702 (2017).
[5] P. Klein, A. Lichtenberger, S. Küchemann, S. Becker, M. Kekule, J. Viiri, C. Baadte, A. Vaterlaus, and J. Kuhn, Visual attention while solving the test of understanding graphs in kinematics: An eye-tracking analysis, Eur. J. Phys. **41**, 025701 (2020).
[6] M. Kekule, Students' approaches when dealing with kinematics graphs explored by eye-tracking research method, in *Proceedings of the 2014 Frontiers in Mathematics and Science Education Research Conference*, pp. 108–117, https://www.scimath.net/download/students-approaches-when-dealing-with-kinematics-graphs-explored-by-eye-tracking-research-method-9632.pdf.
[7] M. Kozhevnikov, M. A. Motes, and M. Hegarty, Spatial visualization in physics problem solving, Cogn. Sci. **31**, 549 (2007).
[8] R. H. Tai, J. F. Loehr, and F. J. Brigham, An exploration of the use of eye-gaze tracking to study problem-solving on standardized science assessments, Int. J. Res. Method Educ. **29**, 185 (2006).
[9] M. J. Tsai, H. T. Hou, M. L. Lai, W. Y. Liu, and F. Y. Yang, Visual attention for solving multiple-choice science problem: An eye-tracking analysis, Comput. Educ. **58**, 375 (2012).
[10] G. Zavala, S. Tejeda, P. Barniol, and R. J. Beichner, Modifying the test of understanding graphs in kinematics, Phys. Rev. Phys. Educ. Res. **13**, 020111 (2017).

[11] A. Susac, M. Planinic, A. Bubic, L. Ivanjek, and M. Palmovic, Student recognition of interference and diffraction patterns: An eye-tracking study, Phys. Rev. Phys. Educ. Res. **16**, 020133 (2020).

[12] A. M. Madsen, A. M. Larson, L. C. Loschky, and N. S. Rebello, Differences in visual attention between those who correctly and incorrectly answer physics problems, Phys. Rev. ST Phys. Educ. Res. **8**, 010122 (2012).

[13] J. Viiri, M. Kekule, J. Isoniemi, and J. Hautala, Eye-tracking the effects of representation on students' problem solving approaches, in *Proceedings of the Annual FMSERA Symposium, 2016* (Finnish Mathematics and Science Education Research Association, 2017), https://jyx.jyu.fi/bitstream/handle/123456789/55879/viirietalfmsera2016.pdf?sequence=1.

[14] P. Klein, S. Küchemann, S. Brückner, O. Zlatkin-Troitschanskaia, and J. Kuhn, Student understanding of graph slope and area under a curve: A replication study comparing first-year physics and economics students, Phys. Rev. Phys. Educ. Res. **15**, 020116 (2019).

[15] J. Viiri, J. Kilpeläinen, M. Kekule, E. Ohno, and J Hautala, Eye-movement study of mechanics problem solving using multimodal options, in *Research and Innovation in Physics Education: Two Sides of the Same Coin* (Springer, Cham, 2020), pp. 145–154.

[16] A. Schüler, Investigating gaze behavior during processing of inconsistent text-picture information: Evidence for text-picture integration, Learn. Instr. **49**, 218 (2017).

[17] A. R. Strohmaier, K. J. MacKay, A. Obersteiner, and K. M. Reiss, Eye-tracking methodology in mathematics education research: A systematic literature review, Educ. Stud. Math. **104**, 147 (2020).

[18] C. Romesburg, *Cluster Analysis for Researchers* (LULU Press, North Carolina, 2004).

[19] P. J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. **20**, 53 (1987).

[20] See https://cran.r-project.org/web/packages/cluster/cluster.pdf.

[21] See https://cran.r-project.org/web/packages/factoextra/index.html.

[22] See https://cran.r-project.org/web/packages/dendextend/index.html.

[23] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.17.013102 for a visualization of dendrogram entanglement.

[24] For more details, see P. E. Downing, Interactions between visual working memory and selective attention, Psychol. Sci. **11**, 467 (2000).