International Journal of Engineering Education Vol. 35, No. 6(A), pp. 1839–1850, 2019
Printed in Great Britain

0949-149X/91 $3.00+0.00

Development and Psychometrics of a Freely Available Mechanical Aptitude Test*

DIANA BAIRAKTAROVA
Virginia Tech, Department of Engineering Education, 367 Goodwin Hall, Blacksburg, VA 24061, USA. E-mail: dibairak@vt.edu

DAVID REEPING
Virginia Tech, Department of Engineering Education, 659 McBryde Hall, 225 Stanger Street, Blacksburg, VA 24061, USA.
E-mail: dreeping@vt.edu

Mechanical aptitude refers to individual differences in understanding and learning how simple machines work. Tests of mechanical aptitude are predictive of performance in engineering jobs and capability to learn about mechanical processes. The advancement of technology has led to existing mechanical aptitude tests becoming dated. Commonly used tests are known to be gender-biased, limited in use (diagnostics tests), and are not freely available for use in educational settings. This work presents the development of a Mechanical Aptitude Test (MAT). The mechanical aptitude items were designed and tested across four phases in large samples of engineering and non-STEM students across four U.S. universities (n = 1,718). An item analysis of the last phase (n = 599) was conducted to screen questions not meeting established criteria for item difficulty and item discrimination. After, a one-factor confirmatory factor analysis was run with diagonalized weighted least squares. The one-factor confirmatory factor analysis fit well with exceptional fit indices (CFI = 0.994, TLI = 0.993, RMSEA = 0.02 90%CI[0.004,0.3], SRMR = 0.059), albeit a rejected model chi-square, $\chi^2(34) = 146.939$, p = 0.042. The current MAT scale consists of 17 multiple choice items, narrowed down from a larger bank of 68 items, covering topics related to mechanical insight, mechanical knowledge, shop geometry and measurement, and tool knowledge.

Keywords: mechanical aptitude; instrument development; item analysis

1. Introduction

Recently in STEM education, much attention has been given to developing and improving students' spatial reasoning, thus emphasizing the value of spatial thinking and success in STEM professions. A construct related to spatial thinking is mechanical aptitude, which refers to individual differences in understanding and learning how simple machines work, how to use and understand tools and machines, and the ability to understand and apply physical and mechanical principles. Mechanical aptitude has historically only been used in engineering education for diagnostic purposes, but recent tests of mechanical aptitude have been reframed as predictors of performance in manufacturing and production jobs. Mechanical aptitude is measured with items that require the application of mechanical principles, identification of objects, or problem-solving. Most popular U.S. measures of mechanical aptitude such as The Wiesen Test of Mechanical Ability (WTMA) and subscales of the Armed Services Vocational Aptitude Battery (ASVAB) were originally developed to serve the United States Army. However, the presentation of the tests has become dated due to advancements in technology. Moreover, commonly used diagnostic tests are known to be gender-biased, have limited use, and are not freely available for use in educational settings.

We have developed a new measure of mechanical aptitude for use in engineering and STEM education and similar contexts. This paper describes the development, item analysis, and confirmatory factor analysis. We then discuss its potential application to engineering education. The new scale has several advantages for educators: (a) the questions are based on common everyday objects and events rather than objects or events encountered primarily or only in academic courses (eg. physics or chemistry); (b) the scale uses topics and objects common in engineering; (c) the scale is useful for many educational applications beyond engineering; and (d) will be freely available to educators and researchers.

2. Background

This section provides a brief introduction to aptitudes and abilities along with a discussion of their cruciality to learning and performance in particular domains. What follows is a comprehensive historical review of measuring mechanical aptitude including what the measure predicts and the types of tests currently used in industrial and educational settings.

2.1 Aptitudes and abilities

Ability and aptitude are often viewed as similar constructs, but the psychological literature makes

* Accepted 18 August 2019.

a distinction between them. Ability is the capacity to perform a particular act or task, either physical or mental [1]. Ability is considered to be an attribute of the individual revealed by performance differences in the levels of task difficulty on a defined class of tasks [2]. The description by Gottfredson [3] is incisive: "Abilities are what people can do, not their style of doing it. Abilities are not the bodies of knowledge that people amass but their aptness in amassing them" (p. 117). The modern understanding of the constitution of abilities has been informed by the conceptualizations of J. P. Guildford and J. Carroll, who are often cited regarding ability testing for career guidance.

Aptitudes are better understood as potential abilities [4]. Aptitudes reflect how likely an individual is to be successful in performing a given task. Arulmani [5] describes an aptitude as "reflect[ing] what one would be naturally good at, the person's talents and capabilities, and the strength of likelihood for achievement in a particular area. If interests are the steam in a locomotive, aptitudes could represent the engine: the actual ability to move toward and be successful in the execution of a specific set of tasks" [5, p. 609]. Aptitudes are predisposing all of "these characteristics (e.g., experience, ability, knowledge, motivation, and regulatory processes) that an individual brings together to perform in a particular situation" [6, p. 79]. Shavelson et al. [6] and Arulmani [5] support the idea that aptitudes are a special kind of individual difference involving the potential to learn or to perform certain tasks.

2.2 Measuring mechanical aptitude

Many engineering students are drawn to engineering because of an interest in learning how things work and creating devices leveraging nature's properties—the underpinnings of mechanical aptitude [7]. Mechanical aptitude is essential to engineering and has a long history in engineering education. Mechanical aptitude encompasses knowledge of topics such as sound and heat conduction, velocity, gravity, and force [8]—all of which are required for common engineering applications. It is typically measured by individuals' performance on a mechanical aptitude test. The test taker is often tasked to recognize which mechanical principle is represented in a situation and apply the principle to a physical problem.

Tests of mechanical aptitude have been developed and used for a variety of purposes. The first test, called the Stenquist Test of Mechanical Aptitude, was created almost a century ago [9]. The Stenquist test includes ten different scenarios where the test taker is asked to observe common images of mechanical objects and determine which image best fits in another set of images. The test attempts to measure the spatial visualization ability of the test taker, which has a strong correlation with "intelligence," as measured by the Army Alpha Examination [10].

The test is used to predict the performance of an individual in specific career fields [9]. For example, Simpson, a psychologist at the Institute for Juvenile Research in Chicago, used the Stenquist test to investigate the relationship between the performance on the test and mechanics' job experience. The study claimed that scores on the Stenquist test were significant predictors of mechanical aptitude and that the Stenquist test could detect improvements in mechanical ability gained through working in a mechanical trade. Simpson [11] suggests that the Stenquist test used in conjunction with an occupational inventory and an intelligence test provides a mechanism for ascertaining mechanical aptitude in adults.

Another test, The McQuarrie Test for Mechanical Ability [12], was created in 1927 to measure an individual's ability to recognize space relation, his or her speed of decision and movement, head and eye coordination, muscular control, and visual acuity. The test is divided into several sections: Tracing, Tapping, Dotting, Copying, Location, Blocks, and Pursuits [12]. The test has been used broadly in educational settings. For example, McQuarrie's test was implemented to measure the engineering aptitude of first-year engineering students at Oregon State University, to select trainees for mechanical occupations in the National Defense program, and to measure the mechanical aptitude of business majors [13]. It is recommended that the McQuarrie test should be used with other tests to maximize its predictive power [13].

The McQuarrie and Stenquist tests are moderately correlated ($r = 0.66$) which indicates that although the tests are different in format, both measure a similar cluster of abilities [14]. However, the group of abilities measured by the McQuarrie and Stenquist tests is distinctly different from the group of abilities measured by other prominent tests, such as The Army Alpha and the Kohs Block Design Test. The early Army Alpha test measured verbal and numerical ability, the ability to follow directions, and knowledge of different principles of construction. The Kohs Block Design, a psychometric performance test, tasks individuals to arrange groups of multicolored blocks and to copy patterns presented on test cards in an attempt to measure IQ. The Stenquist test is more closely associated with the Army Alpha and Kohs measure than the McQuarrie test [15]. The different tests are summarized in Table 1.

Table 1. Early tests of mechanical aptitude and their uses

MA Test	Year of Creation	Measures	Uses
Stenquist Test of Mechanical Ability	1922	Measures mechanical knowledge.	Predict the performance of an individual in a specific career field.
McQuarrie Test for Mechanical Ability	1927	Ability to recognize space relation, speed of decision-making and movement, head and eye coordination, muscular control and visual acuity.	Diagnose the engineering aptitude of first-year engineering students in Oregon State University; selection of trainees for mechanical occupations in National Defense program; measure the mechanical aptitude of business majors.
Army Group Examination Alpha	1936	Verbal and numerical ability, ability to follow directions, and knowledge of different principles of construction.	To determine a soldier's capability of serving, his job classification, and his potential for a leadership position.
Kohs Block Design	1920	Measures the ability of an individual to arrange groups of multi-colored blocks and to copy patterns presented on test cards.	Designed to be an IQ test; later has been administered to school children with exceptionalities.

2.3 Limitations of existing mechanical aptitude tests

Since the first test was created in 1922, close to 30 tests of mechanical aptitude and ability have been developed. The underlying concepts measured by these items include sound and heat conduction, velocity, gravity, and force [8], and most of these tests are predictive of performance in manufacturing/production jobs [8, 9, 12, 14]. More contemporary tests also include questions about the general ability to learn about mechanical principles as a result of everyday living. Over the last 50 years, common measures were developed to serve the United States Army. Tests such as the Wiesen Test of Mechanical Ability (WTMA), Armed Services Vocational Aptitude Battery (ASVAB), and Bennett Mechanical Comprehension Test (BMCT) are improvements over the older tests. Rather than comparing a group of pictures, modern tests involve more solid principle- and text-based questions.

An area of criticism persists concerning items that may be gender-biased in favor of male test-takers. The severity of the bias may be exaggerated, as four different studies comparing the relative adverse impact of the WTMA and BMCT demonstrate that the WTMA has less of an adverse impact on women than the BMCT does [16]. Also, a survey of existing tests reveals a lack of consistency in the facets that are measured. Some tests cover mechanical knowledge, mechanical insights, tool knowledge, and shop arithmetic (e.g., Weisen Test of Mechanical Aptitude), while others measure a broad range of content from general science, mathematical reasoning, mathematical knowledge, word knowledge, verbal expression, electronics information, and assembling objects (e.g., ASVAB). In addition, the advancement of technology leads to the content of the tests becoming dated especially if they reference "current technology" of the time period.

A final, critical issue is that the most commonly used instruments are neither free nor cost-effective for use in academia. Given the cost constraints, mechanical aptitude and ability tests have been mainly used in military and industry, and have been underutilized in education and educational research. For example, the WTMA has been used to determine performance in industrial occupations in sectors such as utility companies, machine operators for a textile manufacturer, public transportation organizations, diesel engine manufacturers, and maintenance. In educational settings, the tests have been used primarily to: (1) predict high school and freshmen college students' vocational interest [13, 17–20]; (2) measure general intelligence [21]; and (3) find relationships between mechanical aptitude and ability and performance on specific discipline-based subjects [22]. More recently, the tests have been used to investigate gender differences in technical abilities [23–25]. The use of the tests for parsing out gender differences is troubling considering that many of the existing tests have been criticized for being gender-biased.

In engineering education, the tests have been utilized primarily in diagnostic studies to predict academic performance [13, 22] or as a college entrance test [26]. There are only three recent studies that use mechanical aptitude or ability in engineering education of which we are aware. The first two studies examine the self-efficacy of female engineering students based on engineering task performance [27, 28]. The studies reveal that male engineering students have more confidence in their engineering ability than female students, despite any gender differences in observed performance. The findings

hold regardless of task difficulty and when students' mechanical aptitude and prior experience with a similar task are controlled. Students' self-evaluations of their ability and their mechanical aptitude were both strong, significant predictors of actual engineering task performance [28]. The third study compares mechanical aptitude, prior experiences, and engineering aptitudes of mechanical female engineering students [23]. The 16-question mechanical insight practice test from a Levy & Levy workbook [7] was used with the publisher's permission. The Levy and Levy test originally served as preparation for civil service, military, and trade exams, and features questions about gears, pipes, linkages, and other mechanisms. Findings from the study reveal that female students rely mainly on coursework to develop technical aptitude. Those with higher scores were more likely to choose to study engineering because they liked "figuring out how things work" [23, p. 269]. The three studies provided evidence of the value of mechanical aptitude and ability measures in research about some of engineering education's most pressing issues and underscore the scarcity of published research that uses them.

3. Development of the mechanical aptitude test

As research in engineering education becomes more sophisticated, there is a clear need for a mechanical aptitude test that can be administered with little to no cost. For example, one of the researchers from this study sought to investigate the effect of the presence of mechanical objects on performance in an engineering assembly task, and it was critical to measure students' mechanical aptitude to make a connection between the units of study [29]. However, there were no affordable mechanical aptitude tests available for use—even with the 40% educational discount offered by the publishers of one particular popular test. Despite efforts to make arrangements with the publisher and the test author, the price far exceeded a reasonable budget for the research and was well beyond the price tag of all but the most well-funded research programs. The price severely limits the use of extant mechanical aptitude tests in research and makes the collection of pilot data almost impossible. Given the expenses to purchase the WTMA test, the first author created a parallel test. It should be noted here that the first author has a background in Mechanical Engineering and Engineering Education, so she was motivated to develop a new scale to be used in both research and engineering pedagogy.

3.1 Construct definition

The proposed dimensionality of the scale was one construct, mechanical aptitude. Here, mechanical aptitude refers to the ability to comprehend and apply the principles of mechanical objects to solve problems [30]. As in most mechanical aptitude tests, the questions in the new test include concepts about simple mechanical objects like levers, pulleys, gears, springs, simple electrical circuits and tools. The underlying concepts measured by the test items do not only ask the test-taker to apply knowledge but to use mechanical insight and intuition. This test was not designed to distinguish between the inherent ability of the test-taker and their retention of previous exposure to these concepts, thereby making the items effect indicators and appropriate for scale development [31].

3.2 Generating items for the instrument

The initial version of the scale consisted of four sections and 68 items. The items across all four sections covered four facets of mechanical aptitude: mechanical insight (MI), mechanical knowledge (MK), tooling knowledge (TK), and shop geometry and measurement (SGM). All four facets were different types of indicators for mechanical aptitude used to generate questions, not necessarily independent constructs. Therefore, unidimensionality was posited. Initial screening with an exploratory factor analysis of the 68 items after the first pilot test, specifically using the scree plot, with 384 first-year engineering students revealed a strong eigenvalue more than double the following eigenvalue—preliminary evidence of unidimensionality.

The first section contained 33 multiple-choice items. Two of these items had yes/no responses and the remaining 31 items had four or five response options. The second and third sections of the exam consisted of 16 pictures of tools or devices (e.g., a voltmeter, calipers) that the respondent was instructed to name (part A) and write a brief description of the use of the tool or device (part B). The fourth section asked three open-ended, short-answer questions about a series of pictures of tools and how they relate to each other (e.g., a screw, a washer, and a bolt).

The mechanical insight problems ask questions based on everyday life. The answers require more logical thinking and observations than any prior knowledge of physics principles. Some questions assessing tool knowledge and ask for recognition of units about current, voltage, pressure, and the ability to read simple diagrams, drawings or geometrical shapes. Moreover, the mechanical knowledge questions ask about basic principles that are introduced in the middle school physics curriculum, relying on Common Core standards. For example, one of the questions shows a sketch of two levers and asks which one is the most efficient. The tools

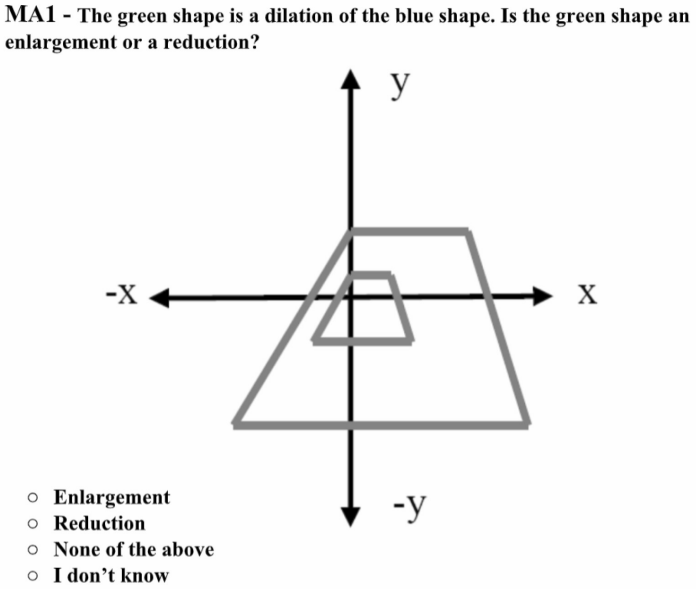

Fig. 1. Example SGM question, green is the larger shape and blue is the smaller.

knowledge and shop geometry and measures questions were created based on experience and empirical data speaking to the value of both to the engineering process. The researchers designed the test deliberately to include the principles mentioned above: tools, engineering practice, and everyday life examples as these are part of the foundational engineering curricula for all undergraduate engineering majors.

For example, Figure 1 shows an example of an SGM question concerning dilation of a shape. Figure 2 shows an MK example question and Figure 3 presents an MI question.

The items about mechanical knowledge and mechanical insight came from examples from middle school physics test bank questions concerning basic principles about sound and heat conduction, velocity, gravity, and force. We have modified the questions and answers of the test bank questions and created images/figures to ensure that the questions used language more appropriate for early college students. The items about shop geometry and measurement and tool knowledge came from one of the researchers' descriptive geometry and manual drawing design education and engineering practice experience. We have created the geometry and shop arithmetic questions from scratch. The open-ended tool questions used photo images from Internet sites with a Creative Common (CC) license.

3.3 Refining items and formatting

Prior to the first administration, seven professional engineers independently evaluated the mechanical aptitude items, assessing the facet of mechanical aptitude that best represented their level of difficulty and consensus on correct answers. Within each facet, items were graded on a difficulty continuum to ensure that moderately difficult, difficult, and some very difficult items were included. The coefficient of agreement, Fleiss' Kappa, of the facets was found to be 0.84 and the coefficient of agreement for correct answers was 0.93. Values of 0.75 typically

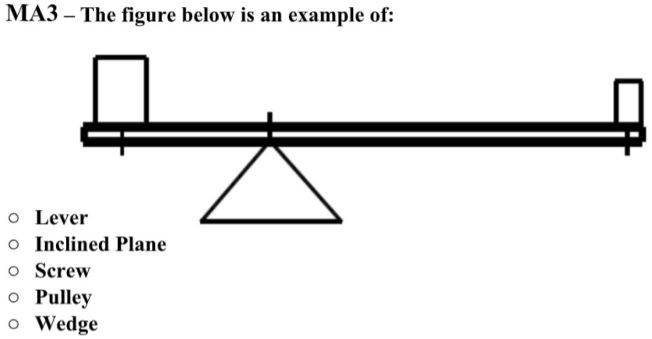

Fig. 2. Example MK question.

MA17 - The fraction 3/8 expressed as a decimal is:

- ○ **0.125**
- ○ **0.250**
- ○ **0.333**
- ○ **0.375**
- ○ **0.425**

Fig. 3. Example MI question.

indicate good agreement among raters but values greater than 0.90 are preferred [32].

Test development involved several iterations with student samples from four universities: Phase I involved initial piloting of the first version of the instrument (V1) with undergraduate engineering students. Phase II built upon lessons learned from Phase I by creating the next version (V2) and testing it with undergraduate engineering students from a different university and non-engineering majors. Phase III was conducted for further refinement and testing of the revision (V3) with undergraduate engineering students from a third university. Phase IV was the culmination of the previous four phases by testing the instrument's fourth iteration (V4) with a large sample of first-year engineering students. The demographics for all four samples are provided in Table 2. Note the final sample is the focus of this paper.

The initial version (V1) of the test was administered to 384 first-year engineering students enrolled in an introductory engineering course at a large university in the Midwestern United States. Students across four class sections completed the instrument as part of a larger study. Most of the students were Caucasian males, and 76 percent of participants reported English as their first language. The sample yielded complete data for 378 students. The multiple-choice items were scored automatically, and the open-ended items were scored by hand using a set of possible answers generated by practicing engineers as the scoring criteria.

There were 51 items in V2 of the test. The open-ended name and function items were retained and the weakest performing multiple-choice questions were removed from V1 based on preliminary item analyses. Of the remaining items, thirty-two questions were multiple-choice items. Another eight were two-part free-response items which displayed color images of tools or devices (e.g., a voltmeter, calipers) that the respondent was instructed to name (part A) and write a brief description of the use of the tool or device (part B). There were also three open-ended, short-answer questions about a series of images of tools and their relationship with one another (e.g., a screw, a washer, and a bolt).

The V2 test was administered as an online survey during class time and students were instructed not to look for the images' names online. The mechanical aptitude items from both scales were presented together in random order. The V2 test was administered to 566 students, engineering ($n = 257$) and non-engineering ($n = 309$), across two institutions to examine how well the test distinguished between STEM and non-STEM samples.

The third phase of the MAT development involved administering the 44 best performing

Table 2. Demographics for Phases I through IV

	Phase I	**Phase II**		**Phase III**	**Phase IV[1]**
N	**384**	**257**	**309**	**169**	**599**
Major	**First-Year Engineering**	**Engineering**	**Non-STEM**	**Engineering**	**First-Year Engineering**
Age (yrs)					
Mean	18.22	21.34	20.91	20.02	18.46
Range	17–21	18–55	18–56	18–71	18–28
Sex (%)					
Male	82.8	69.3	13.6	78.7	55.3
Female	17.2	24.5	85.1	17.8	34.7
No response					10.0
Ethnicity (%)					
African-American/Black	1.8	5.8	2.9	4.1	4.2
Asian or Pacific Islander	23.8	9.7	11.0	19.5	22.2
Hispanic	2.1	9.3	43.4	4.7	3.7
White	60.6	56.0	29.1	54.0	55.8
Other[2]	11.8	12.9	13.0	7.7	10.4
Prefer not to answer					19.7
No response					12.4
Language (%)					
English as first language	76	79.0	63.1	76.3	72.8

Note: (1) Phase IV is the focus of this paper, (2) "other" includes Native American and multiple ethnicities.

items from the first two phases of development to another sample of 169 engineering students. Eight mechanical insight, 10 mechanical knowledge, 10 shop geometry and measurement, and 16 open-ended tool name and function items were used.

Finally, the phase reported in this paper, phase IV, took the lessons learned from the previous phases and looked specifically at the remaining 22 multiple-choice items as a single factor, mechanical aptitude. Open-ended items were not considered because the factor structure of the main multiple-choice items was of concern to this work and require more sophisticated statistical techniques to incorporate.

3.4 Item analysis

We used classical test theory to evaluate the MAT in Phase IV. Accordingly, the item difficulty, item discrimination, mean inter-item correlation, and overall alpha were calculated. Table 3 provides a summary of the item analysis for the MAT.

Robinson et al. [33] advocates for an overall alpha greater than 0.8 and mean inter-item correlation greater than 0.3 while Clark and Watson [34] suggest an inter-item correlation between 0.15 and 0.5 across constructs. McCoach, Gable and Madura [35] claim researchers often consider an overall alpha of 0.7 to be acceptable for most research purposes. The MAT had an overall alpha of 0.857 and mean inter-item correlation of 0.235, falling within the recommended intervals and just below Robinson et al.'s [33] recommendation for inter-item correlation.

The reliability could be hampered by certain items, which is found by examining the alpha-if-deleted column. Items with alphas-if-deleted greater than the overall alpha indicating they are troubling items impacting the overall reliability of the instrument in a negative way and should be removed. Reviewing the item analysis revealed that the items MAT.1, MAT.2, and MAT.4 had values of alpha-if-deleted greater than the overall alpha, meaning the three could be candidates for omission in future applications.

Moreover, the items needed to be screened for difficulty and discrimination. Jorion et al.'s [36] standards for "excellent" were applied as the boundaries on difficulty and discrimination. The minimum discrimination, discrim_min, was set to 0.2 and the acceptable interval for item difficulty was set to 0.2–0.8, diff_min and diff_max respectively. Items not within the pre-set standards were considered candidates for removal from the overall instrument. Figure 4 displays the item difficulty and item discrimination with the boundaries appended to the plot. Appropriate items are denoted by points in the top middle box of Figure 4.

The item analysis revealed items not meeting the recommended thresholds for item discrimination and item difficulty. Four items fell outside of the difficulty range (MAT.19, MAT.20, MAT.21, MAT.22) while item MAT.1 did not meet the minimum level of item discrimination. The five items were removed because the last items, MAT.19–MAT.22, were too easy and MAT.1 did

Table 3. Item Analysis Results

Item*	Mean	Standard Deviation	Item Difficulty	Item Discrimination	α if Deleted
MAT.1	0.23	0.42	0.23	0.049	0.864
MAT.2	0.38	0.49	0.38	0.241	0.859
MAT.3	0.51	0.5	0.51	0.318	0.856
MAT.4	0.55	0.5	0.55	0.215	0.86
MAT.5	0.58	0.49	0.58	0.362	0.854
MAT.6	0.61	0.49	0.61	0.39	0.853
MAT.7	0.64	0.48	0.64	0.397	0.852
MAT.8	0.65	0.48	0.65	0.46	0.85
MAT.9	0.65	0.48	0.65	0.384	0.853
MAT.10	0.66	0.47	0.66	0.469	0.85
MAT.11	0.66	0.47	0.66	0.412	0.852
MAT.12	0.67	0.47	0.67	0.471	0.85
MAT.13	0.67	0.47	0.67	0.414	0.852
MAT.14	0.67	0.47	0.67	0.446	0.851
MAT.15	0.68	0.47	0.68	0.42	0.852
MAT.16	0.71	0.45	0.71	0.544	0.847
MAT.17	0.74	0.44	0.74	0.505	0.848
MAT.18	0.78	0.42	0.78	0.558	0.847
MAT.19	0.87	0.34	0.87	0.701	0.844
MAT.20	0.87	0.33	0.87	0.657	0.845
MAT.21	0.89	0.32	0.89	0.687	0.845
MAT.22	0.9	0.29	0.9	0.78	0.844

* Items were renumbered for this work in order of item difficulty for ease of reading.

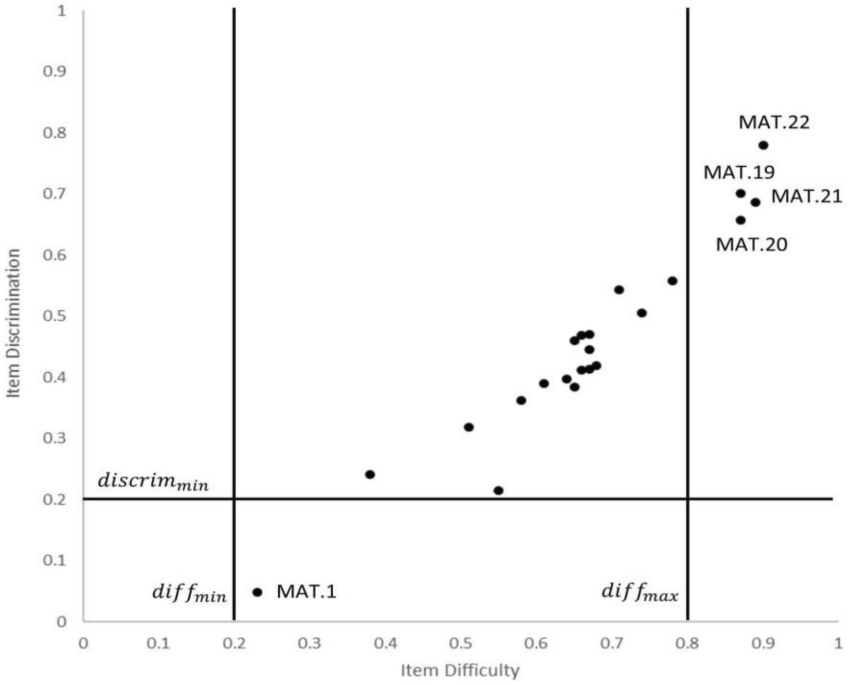

Fig. 4. Item difficulty and item discrimination for the MAT. Recommended levels of difficulty and discrimination are appended as boundaries.

not help in discerning between high and low scoring test-takers.

Structural properties prior to the confirmatory factor analysis could be assessed using a tetrachoric correlation matrix. The tetrachoric correlation is more appropriate for the MAT because the items are dichotomous, not continuous—as assumed by the Pearson correlation coefficient. Figure 5 displays a heat map of the correlations by item. Darker squares indicate stronger correlations. The correlations ranged from 0.107 to 0.631 excluding items with extraordinary low correlations with a specific item including correlations like r_{tet}(MAT.5, MAT.12) $\sim$ 0, r_{tet}(MAT.10,MAT.17) = 0.090, and

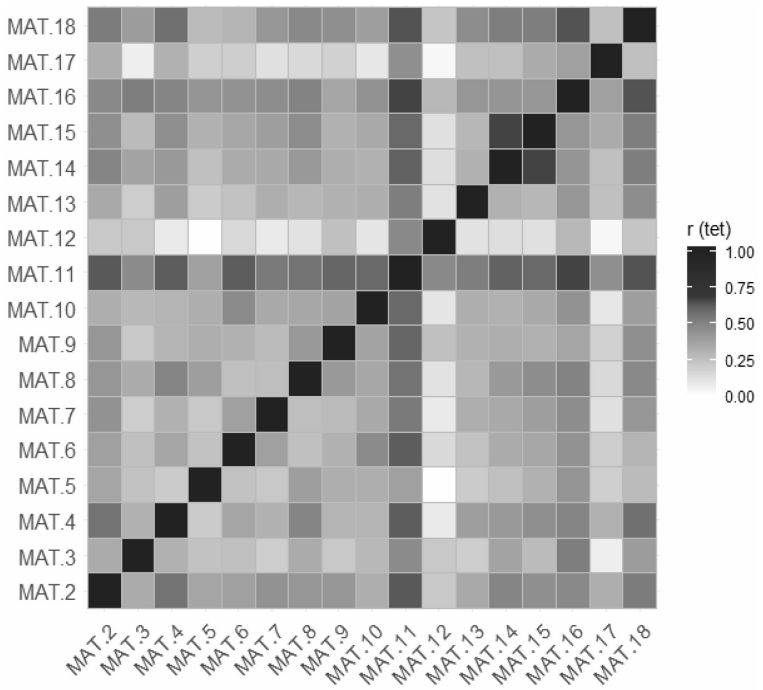

Fig. 5. Inter-item tetrachoric correlation heat map for MAT items. Darker squares indicate a stronger correlation.

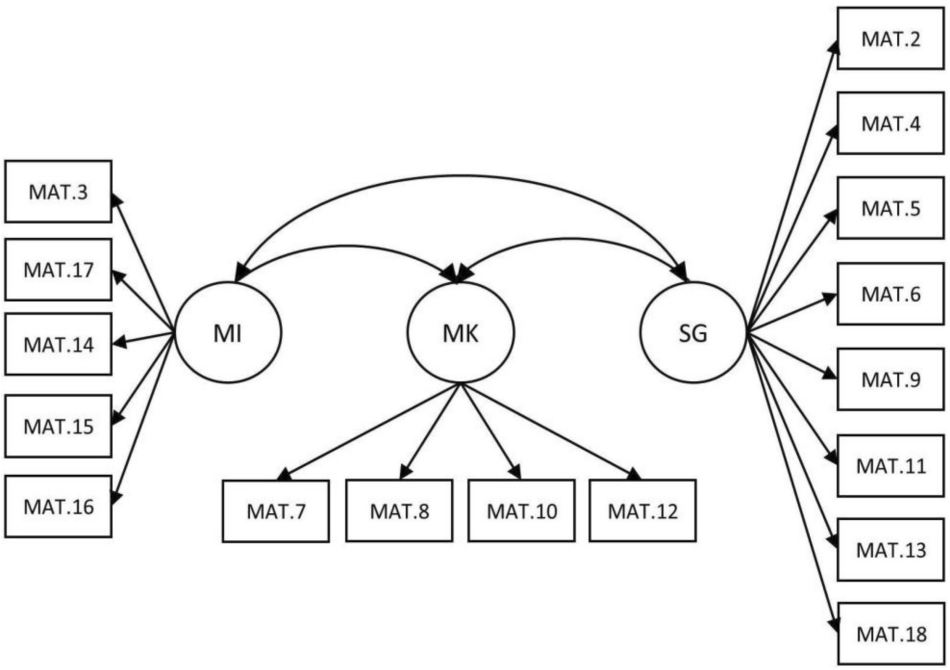

Fig. 6. Three-factor model specification. Note that MI is Mechanical Insight, MK is Mechanical Knowledge, and SG is Shop Geometry and Measurement.

r_{tet}(MAT.3, MAT.17) = 0.058. All items, except MAT.12 and MAT.17, have a visually uniform correlation pattern across the heat map. The uniform correlation provides some evidence of a simple underlying one factor structure.

3.5 Confirmatory factor analysis

The confirmatory factor analysis was conducted in R using the lavaan package [37] with a hypothesized one-factor solution and three-factor solution. The estimator used was Diagonally Weighted Least Squares with a tetrachoric correlation matrix as the input because all the items are binary [38,39]. The models were identified using the unit variance identification constraint [see 40] to avoid needing to constrain a reference variable. The three-factor solution was formed by loading the items with their original content area in the first phase: Shop Geometry and Measurement, Mechanical Knowledge, and Mechanical Insight (Fig. 6). Tool Knowledge was not included in the factor structure because the open-ended questions were not analyzed as part of the model.

CFA models require a large sample size, and implementations of structural equation modeling (SEM) techniques like CFA generally fall below the necessary sample size to draw sound conclusions. For instance, Westland's [41] review of 74 SEM studies in four management and information systems journals revealed an average sample size of 375—50 percent of the minimum sample size needed to support the authors' claims. The simplicity of the model here is somewhat insulated from the discussion in the SEM literature since the model is a more routine CFA. The 599 total observations from a single sample were deemed to be well above the typical sample size requirements.

The results of the CFA for the one-factor model are given in Table 4. The three-factor model was attempted, but R returned a warning about a non-positive definite matrix and negative variance estimates. Further inspection of the covariance matrix revealed a high correlation between the three proposed factors, which can cause non-positive definite matrices and negative variance estimates. Shop Geometry and Measurement strongly correlated with Mechanical Knowledge (r_{tet} = 0.848), while Mechanical Insight and Mechanical Knowledge (r_{tet} = 0.963) and Mechanical Insight and Shop Geometry and Measurement (r_{tet} = 0.999) correlated nearly to the point of being indistinguishable factors from one another. Such high correlations led the researchers to accept the one-factor solution as the preliminary statistical model for the multiple-choice questions in the MAT.

The one-factor model presented with excellent fit, albeit a rejected model chi-square at the typical α = 0.05 level, $\chi^2(34) = 146.939$, p = 0.042. All other fit indices were within the appropriate bound as recommended by Jorion et al. [36] and Kline [40]. The range of the pattern coefficients was 0.247 to 0.787 with standard errors no bigger than 0.027. All pattern coefficients were significant at the α = 0.001 level. The pattern coefficients are displayed in Table

Table 4. Summary of Confirmatory Factor Analysis Results

Index	Recommended Value[2]	One Factor Solution
df		34
Model χ^2	Low relative to *df*	146.939*
Baseline Indices		
CFI	> 0.95	Good (0.994)
TLI	> 0.95	Good (0.993)
Population Error		
RMSEA	< 0.10	Excellent (0.020)
RMSEA 90%CI		[0.004, 0.030]
SRMR	< 0.10[1]	Good (0.059)

* Significant at $\alpha = 0.05$, [1] Value recommended by Kline (2016), [2] Recommended values from Jorion et al. [35].

Table 5. Pattern Coefficient Estimates

Item	Patten Coefficient Estimate	Standard Error	*z*-value
MAT.2	0.362	0.027	13.173
MAT.3	0.463	0.027	17.408
MAT.4	0.247	0.027	9.066
MAT.5	0.501	0.026	19.017
MAT.6	0.540	0.026	20.519
MAT.7	0.539	0.027	20.238
MAT.8	0.671	0.026	26.133
MAT.9	0.528	0.027	19.798
MAT.10	0.687	0.026	26.642
MAT.11	0.554	0.027	20.691
MAT.12	0.667	0.026	25.813
MAT.13	0.546	0.027	20.332
MAT.14	0.633	0.026	24.164
MAT.15	0.567	0.027	21.080
MAT.16	0.765	0.026	29.773
MAT.17	0.699	0.027	26.107
MAT.18	0.787	0.027	29.104

5. Equating the pattern coefficients in a test for weak invariance between genders resulted in a non-ignorable decrease in the CFI greater than 0.1. The decrease implies the measurement may be structurally different between genders. Evidence from the literature may suggest the lack of weak invariance is related to women underestimating their confidence in performing tasks associated with mechanical aptitude [27,28].

4. Discussion and implications

This work presented the preliminary structure for the MAT, including a classical test theory approach to evaluating the items in the most recent version of the instrument. The instrument was designed to measure the construct of mechanical aptitude using four different types of indicators. Examining the interitem tetrachoric correlations and results for screening the number of factors to extract in previous samples provided evidence of an underlying one-factor structure. Attempts to estimate alternative confirmatory factor models containing more than one factor revealed nonignorable correlations among factors. Certainly factors are allowed to correlate, hence the use of oblique rotations in exploratory factor analyses as opposed to imposing strict orthogonality. However, such models are not admissible when factors correlate to the extent seen in the intended three-factor model. Aggregating the factors was judged to be the soundest decision. The results of the CFA provide evidence for a strong one-factor solution with exceptional fit indices.

The MAT is offered to educators for diagnostic purposes and creating low and high structured learning environments for students with high mechanical ability and students with low mechanical ability. The test can be used for no cost in educational settings. The creators of MAT plan for educators to have free access to the test banks online. When students with low mechanical aptitude are identified, instructors will receive a report with feedback highlighting the missed questions along with suggested topics, activities, and instructional models that can be suggested to students to enhance their mechanical aptitude. Lastly, the model will include a link to a virtual tool library. The model will be highly structured to address the fact that students with low ability do better with highly structured environments. The test can be used in the beginning of a course as a means to understand students' misconceptions and struggles with particular topics.

The measurement invariance between genders is disappointing but provides a discussion point for measuring mechanical aptitude. Other tests have been shown to gender-biased, but the development of a test specifically defined to be free of the bias still presents with issues. What the difficulties might be indicative of is a problem in how mechanical aptitude is conceptualized, placing the issues not in the items but in the construct definition. What is meant by mechanical aptitude might be biased toward males in its conceptual formulation, which naturally follows into the item generation stage of instrument development. Alternatively, other mediating variables may need to be incorporated to account for what is known about women underestimating their confidence in performing tasks associated with mechanical aptitude [27, 28]. Such an analysis would go beyond conventional analyses of measurement invariance. Future work could serve to explore the link in gender-bias between mechanical aptitude's construct definition and its items. This lack of measurement invariance does not diminish the practical utility of the instrument presented here, as its latent structure was found to be strong across its administrations. Care must be taken in comparing scores between genders, however.

The instrument is also up-to-date. The MAT uses solid principle- and text-based questions as opposed to the unclear pictorial exercise in older tests. In contrast to the most commonly used and validated mechanical aptitude tests are still on paper and pencil (workbooks), with black and white images, and in 2-D environment, the MAT will be taken exclusively online.

5. Conclusion

Literature shows that mechanical aptitude is a crucial attribute for engineering students; yet, mechanical aptitude tests have been underutilized in education. Some of the content in these tests is becoming dated as technology advances, many commonly used tests are known to be gender-biased, and not freely available for use in educational settings. This MAT scale remedies some of the cons of the existing mechanical aptitude tests as the new scale is cost-effective and freely available measure. The scale is up-to-date and easy to administer—providing more applications in engineering education beyond diagnostics. The gender-biased component is an avenue for future work to determine where the latent differences contribute to the difficulty in measurement.

In this paper we introduced the measure, summarized its development and preliminary psychometric analysis, and discussed its potential application to engineering education. Along with the main focus of this paper to introduce the new scale, the researchers hope to engage the community in a meaningful discussion of the MAT potential application to engineering education and research in engineering education. The role of engineering educators is to train future engineers to perform in the field, and the authors of this paper hope that this work will encourage further discussion and community feedback to the suggested here methods to help increase or enhance engineering students' mechanical aptitude.

References

1. R. E. Snow, Aptitude, learner control, and adaptive instruction, *Educational Psychologist*, **15**(3), pp. 151–158, 1980.
2. J. B. Carroll, *Human cognitive abilities: A survey of factor-analytic studies*, Cambridge University Press, 1993.
3. L. S. Gottfredson, The challenge and promise of cognitive career assessment, *Journal of Career Assessment*, **11**(2), pp. 115–135, 2003.
4. R. E. Snow, Aptitude theory: Yesterday, today, and tomorrow, *Educational Psychologist*, **27**(1), pp. 5–32, 1992.
5. G. Arulmani, Assessment of interest and aptitude: A methodologically integrated approach, In *Handbook of Career Development*, Springer, New York, NY. pp. 609–629, 2013.
6. R. J. Shavelson, R. W. Roeser, H. Kupermintz, S. Lau, C. Ayala, A. Haydel, S. Schultz, L. Gallagher and G. Quihuis, Richard E. Snow's remaking of the concept of aptitude and multidimensional test validity: Introduction to the special issue, *Educational Assessment*, **8**(2), pp. 77–99, 2002.
7. J. U. Levy and N. Levy, *ARCO Mechanical Aptitude & Spatial Relations Tests*, Thomson/Peterson's, 2004.
8. P. M. Muchinsky, Validation of intelligence and mechanical aptitude tests in selecting employees for manufacturing jobs, *Journal of Business and Psychology*, **7**(4), pp. 373–382, 1993.
9. J. L. Stenquist, *Stenquist mechanical aptitude tests*, World Book Company, 1992.
10. D. Brandwein, Army Alpha Intelligence Test. In: Goldstein S., Naglieri J.A. (eds) *Encyclopedia of Child Behavior and Development*, 2011, Springer, Boston, MA.
11. R. M. Simpson, The mechanical aptitudes of 312 prisoners, *Journal of Applied Psychology*, **16**(5), 1932, pp. 485–496.
12. T. W. McQuarrie, A mechanical ability test, *Journal of Personnel Research*, **5**, 1927, pp. 329–33.
13. G. W. Holcomb and H. R. Laslett. A prognostic study of engineering aptitude, *Journal of Applied Psychology*, **16**(2), p. 107, 1932.
14. J. P. Wiesen, WTMA: The Wiesen Test of Mechanical aptitude (version 3.12): Newton, MA: *Applied Personnel Research*, 1997.
15. M. L. Stein, A trial with criteria of the MacQuarrie Test of Mechanical Ability, *Journal of Applied Psychology*, **11**(5), p. 391, 1927.
16. Criteria Corps, Wiesen test of Mechanical aptitude, Retrieved from http://www.criteriacorp.com/solution/wtma. php, 2014.
17. B. Balinsky and C. Hujsa, Performance of college students on a mechanical knowledge test, *Journal of Applied Psychology*, **38**(2), p.111, 1954.
18. C. L. Cooper, Mechanical aptitude and school achievement of negro boys, *Journal of Applied Psychology*, **20**(6), pp.751–760, 1936.
19. L. J. Cantoni, High school tests and measurements as predictors of occupational status, *Journal of Applied Psychology*, **39**(4), p. 253, 1955.
20. L. J. Cronbach and R. E. Snow, *Aptitudes and instructional methods: A handbook for research on interactions*, Irvington, 1977.
21. K. C. Garrison, The use of psychological tests in the selection of student nurses, *Journal of Applied Psychology*, **23**(4), pp. 461–472, 1939.
22. I. T. Littleton, Prediction in auto trade courses, *Journal of Applied Psychology*, **36**(1), pp. 15–19, 1952.
23. A. L. Pereira and M. H. Miller, Gender comparisons of mechanical aptitude, prior experiences, and engineering attitudes for mechanical engineering students, *Journal of Women and Minorities in Science and Engineering*, **18**(3), 2012.
24. L. G. Portenier, Mechanical aptitudes of university women, *Journal of Applied Psychology*, **29**(6), pp. 477–482, 1945.
25. F. L. Schmidt, A theory of sex differences in technical aptitude and some supporting evidence. *Perspectives on Psychological Science*, **6**(6), pp. 560–573, 2011.
26. H. B. Reed. The place of the Bernreuter Personality, Stenquist Mechanical Aptitude, and Thurstone Vocational Interest Tests in college entrance tests, *Journal of Applied Psychology*, **25**(5), 1941, pp. 528–534.
27. S. V. Paunonen and R. Y. Hong, Self-Efficacy and the Prediction of Domain-Specific Cognitive Abilities, *Journal of Personality*, **78**(1), pp. 339–360, 2010.
28. A. Woodcock & D. Bairaktarova, Gender Differences in Self-Evaluation of Students' Performance on Engineering Task, *Journal of Women and Minorities in Science and Engineering*, **21**(3), pp. 255–269, 2015.
29. D. Bairaktarova, W. Graziano and M. Cox, Enhancing engineering students' performance on design task: The Box of Parts, *Journal of Mechanical Design*, 2017.
30. C. Cohen and D. Bairaktarova, A Cognitive Approach to Spatial Visualization Assessment for First-year Engineering Students, *Journal of Engineering Design Graphics*, **82**(3), 2018.
31. R. G. Netemeyer, W. O. Bearden and S. Sharma, *Scaling procedures: Issues and applications*, Sage Publications, 2003.
32. MiniTab Inc. Kappa statistics for Attribute Agreement Analysis. Retrieved from https://support.minitab.com/en-us/minitab/18/help-and-how-to/quality-and-process-improvement/

measurement-system-analysis/how-to/attribute-agreement-analysis/attribute-agreement-analysis/interpret-the-results/all-statistics-and-graphs/kappa-statistics/ Accessed Feb 2019.
33. J. P. Robinson, P. R. Shaver and L. S. Wrightsman. Criteria for scale selection and evaluation, *Measures of Personality and Social Psychological Attitudes*, **1**(3), pp. 1–16, 1991.
34. L. A. Clark and D. Watson, Constructing validity: Basic issues in objective scale development, *Psychological Assessment*, **7**(3), p. 309.
35. D. B. McCoach, R. K. Gable and J. P. Madura, Instrument development in the affective domain, *New York, NY: Springer*, **10**, pp. 978–971, 2013.
36. N. Jorion, B. D. Gane, K. James, L. Schroeder, L. V. DiBello and J. W. Pellegrino, An analytic framework for evaluating the validity of concept inventory claims, *Journal of Engineering Education*, **104**(4), pp. 454–496, 2015.
37. Y. Rosseel, Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA), *Journal of Statistical Software*, **48**(2), pp. 1–36, 2012.
38. R. E. Schumacker and S. T. Beyerlein, Confirmatory factor analysis with different correlation types and estimation methods, *Structural Equation Modeling*, **7**(4), pp. 629–636, 2000.
39. C. H. Li, Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares, *Behavior Research Methods*, **48**(3), pp. 936–949, 2016.
40. R. B. Kline, *Principles and practice of structural equation modeling*, Guilford Publications. 2015.
41. J. C. Westland, Lower bounds on sample size in structural equation modeling, *Electronic Commerce Research and Applications*, **9**(6), pp. 476–487, 2010.

Diana Bairaktarova, PhD, is an Assistant Professor in the Department of Engineering Education and Affiliate Faculty in the Department of Mechanical Engineering at Virginia Tech. Her research explores (a) impacts of concrete and virtual objects on student learning and abilities (mechanical reasoning, spatial and creative abilities), and (b) effective teaching models of professional ethics and empathic design.

David Reeping is a PhD Candidate in Engineering Education at Virginia Tech and a National Science Foundation Graduate Research Fellow. He received his B.S. in Engineering Education with a Mathematics minor from Ohio Northern University. His main research interests include transfer student information asymmetries, agent-based modeling of educational systems, and advancing quantitative and fully integrated mixed methods.