Validation Study for the MirMe® Assessment for 21st Century Skills: Measuring Situational Intelligence (SQ)

MirMe®: online psychometric assessment of 21st century skills

The development of 21st century skills has been identified as a central concern by both ministries of education and employers. These are the skills sought for during college admissions and job interviews, and they significantly influence job fit and future career advancement. They are also instrumental to academic success.[1] Consequently, the development of an objective, rigorous, accurate, and reliable means of assessing 21st century skills is greatly to be desired. This paper describes the validation process for the *MirMe® Assessment System for 21st Century Skills* ("MirMe") created by LogicMills Learning Centre Pte Ltd. Among the 21st century skills measured by MirMe is *Situational Intelligence* ("SQ"), which is assessed via the *Decider* module of the MirMe system. The Decider module within MirMe is the focus of the present study.

There are numerous competing popular schemata for categorizing and populating the list of 21st century skills.[2] In practice, however, mapping between the better-known schemata

1 A suggestive case is documented in the Singapore Ministry of Education Report: "Explicit Teaching of analytical thinking skills (ATS) through games-based facilitation for all courses (in Primary and Secondary schools) for Higher academic achievement." Ministry of Education, Singapore Innovation Fund $1.09m grant research project, completed September 2010. This report describes the validation process for the LogicMills ATS® curriculum involving more than 2,000 students over two years. The LogicMills curricula, delivered in less than 30 hours over a single academic year, was shown to boost scores on high-stakes exams by 16.8% or more ($p < .000$, $r^2 \approx .85$). Students who represent Singapore on the PISA exam (administered by the OECD) typically go through either the LogicMills program or a white-labelled version of it. The influence of LogicMills is perhaps best seen in Singapore's performance on the special tests of the PISA. In 2012, Singapore students came 1st for problem solving skills; in 2015, 1st for collaborative problem solving; and in 2018, 1st in global competencies.

2 Examples of well-known schemes include: the Partnership for 21st Century Learning's "P21", https://www.battelleforkids.org/networks/p21; the framework by UNESCO's International Bureau of Education, http://www.ibe.unesco.org/en/glossary-curriculum-terminology/t/twenty-first-century-skills; ACARA's "General Capabilites," https://www.australiancurriculum.edu.au/f-10-curriculum/general-capabilities/; Singapore's "21st Century Competencies," https://www.moe.gov.sg/education-in-sg/21st-century-competencies; and the 10 skills and four categories of the Assessment and Teaching of 21st Century Skills research group, http://www.atc21s.org/. For a discussion of issues arising from the diverse accounts of 21st century skills, see: C. Joynes, S. Rossignoli & E. F. Amonoo-Kuofi (2019), "21st Century Skills: Evidence of issues in definition, demand and delivery for development contexts," Education Development

is fairly straightforward and which scheme one adopts is largely a matter of convenience. Inter-translation can often turn on a simple matter of nomenclature: *critical thinking skills* are sometimes called *analytical thinking skills* (as, e.g., in Singapore).[3]

The definition of 21st century skills adopted for MirMe was constructed by surveying accounts of 21st century skills articulated in multiple international educational jurisdictions, with particular attention being paid to Singapore, Australia, the UK, and the USA. Additionally, as these skills are often described as "employability skills," particular consideration was also devoted to published and first-hand accounts of 21st century skills by employers. In the authors' experience, the framework articulated below captures between 90% to 95% of 21st century skills described by various ministries and departments of education globally.[4]

The approach adopted for MirMe divides 21st century skills into three major categories:

- *Situational Intelligence* (SQ) — the skills (and concepts) needed to make good decisions in a changing world;
- *Collaborative Intelligence* (CQ) — the skills (and concepts) needed to work with other people; and,
- *Global Intelligence* (GQ) — the skills (and concepts) and qualities needed to be an effective global citizen.

A graphical representation of the relations among these three components is given in Figure 1 below. SQ and CQ may be considered as pillar skills supporting GQ. That is, GQ takes up and integrates the lower pillar skills upon which it rests, resulting in new skills. A useful comparison is the relationship between skill in negotiation (a high-level, integrated skill akin to GQ) and its pillar skills (akin to SQ and CQ). Skill in negotiation depends upon

---

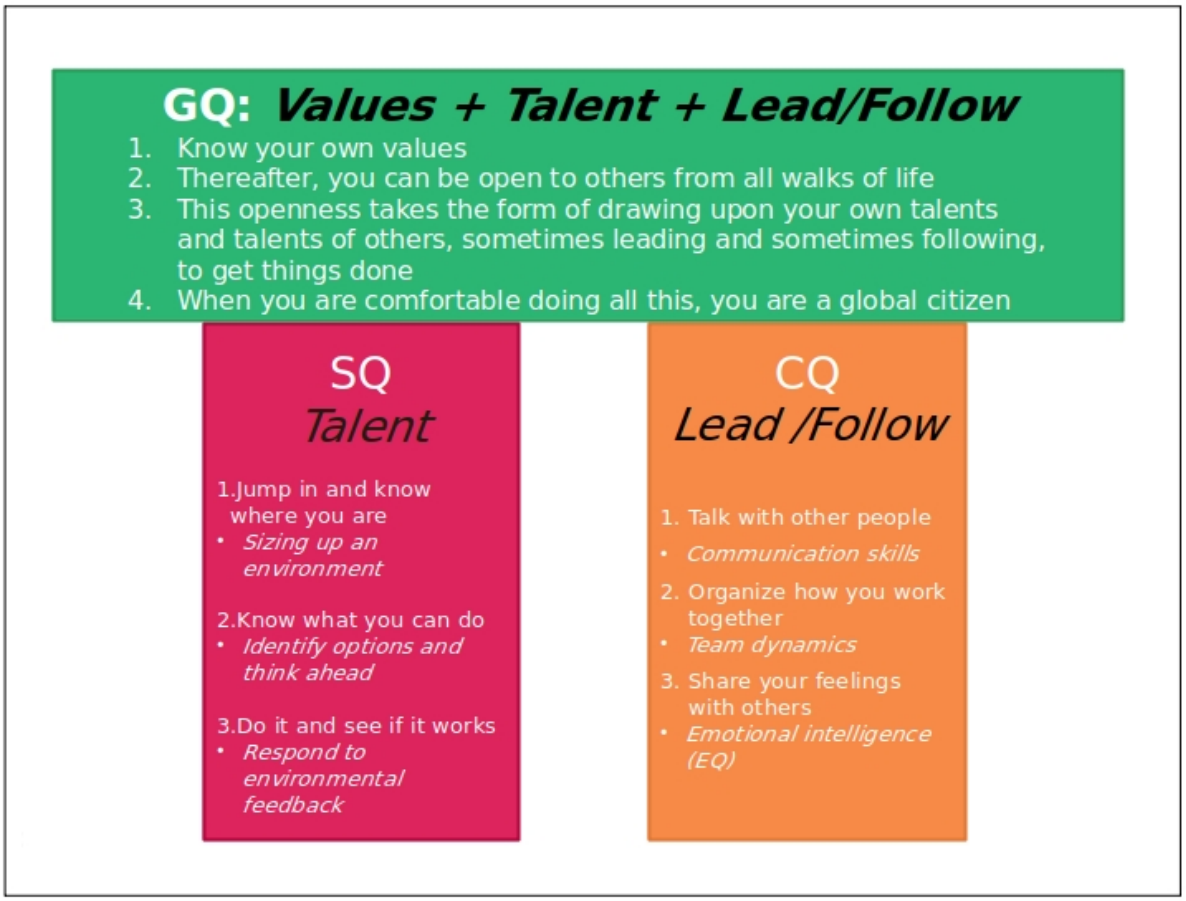
Trust Institute of Development Studies, available at: https://assets.publishing.service.gov.uk/media/5d71187ce5274a097c07b985/21st_century.pdf All sites accessed 31 August 2021.

3 This is, perhaps, due to cultural norms; in Asia direct conflict is usually avoided and it is considered impolite to criticize. With respect to content, though, the terms are identical.

4 The lion's share of the missing 5-10% is typically due to inclusion of country-specific factual knowledge. Singapore, for instance, includes a working familiarity with its history and political institutions in its 21st century skills. Other mapping challenges arise due to diverse notions of digital literacy, which in some cases can be highly specific: e.g., coding is sometimes considered essential, sometimes not; fluency in using specific search engines is sometimes required, other times not.

but is not reducible to its various pillar skills, which include communication skills, skills for setting and prioritising goals, skill in cost/benefit analysis, and so on.

Figure 1. Framework for 21st century skills adopted for MirMe.



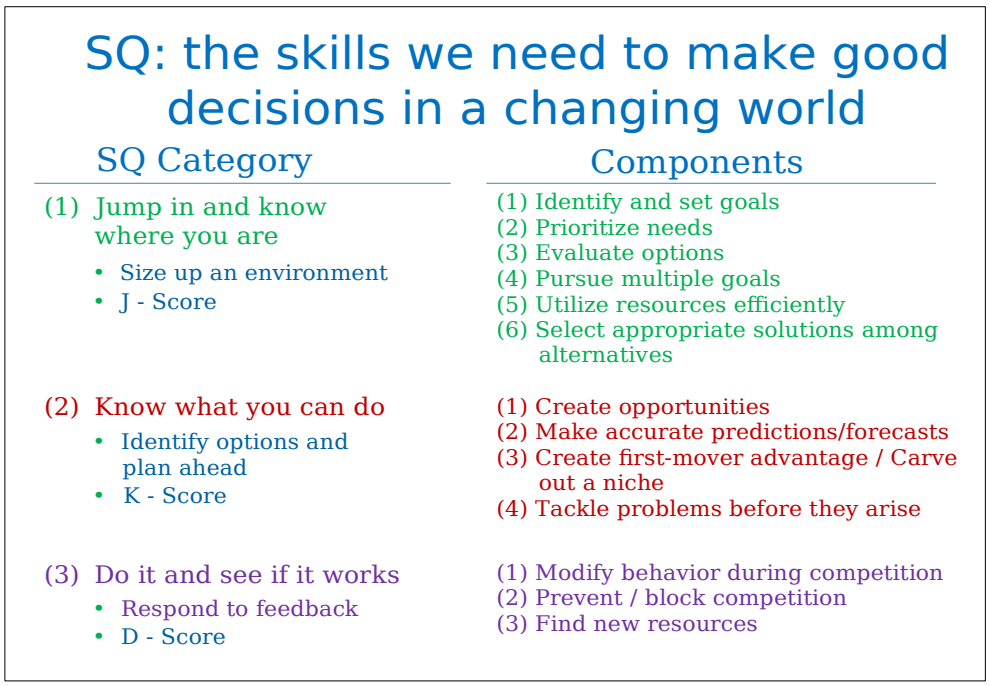
Validation overview

The SQ measures assessed by the Decider module were carefully developed, reflecting insights gathered from more than eight years of observation of how participants played a similar game used in the Decider and how it functioned as an assessment tool in concrete teaching practice. Based on our experience teaching and assessing 21st century skills for over 80,000 Singapore students of various ages, the selected measures all had high face validity. At the time, the Decider had more than 22,000 gameplays and 7,000 users as part of its various validation studies. Some of these studies, especially those involving industry partners, are subject to confidentiality agreements. What follows are studies approved for external publication.

The Decider assessment module consists of an abstract strategy game that is played three times as well as a brief 15 yes/no question survey. Upon completion of the assessment, the Decider calculates an overall summative or holistic score for SQ. This summative score is

calculated from three major components (called "SQ Categories"), each of which in turn is calculated from multiple sub-components (each of which again depends on multiple sub-sub-measures). See Figure 2 below.

Figure 2. Component measures of the Decider SQ assessment.



As may be seen from Figure 2, the Decider is designed to measure SQ, understood as *the skills we need to make good decisions in a changing world*. The Decider's overall *SQ-Score* represents a holistic measure of this.

The SQ-Score is, in turn, built upon three components. In general, a person possessing situational intelligence displays three characteristics: First, you must be able to *Jump in and know where you are* ("J-Score"). That is, one must be able to size up an environment and quickly identify significant patterns and key drivers in that environment. Second, you should *Know what you can do* ("K-Score"). That is, one should be able to identify options and think ahead. Third, you must be able to *Do it and see if it works* ("D-Score"). This means that one is able to assimilate feedback from the environment and adapt to that feedback in a flexible manner. The overall SQ-score is determined by the participant's J-Score, K-Score, and D-Score.

As is usual in the development of a novel psychometric instrument, considerable refinement from a more complex initial form was required. In the present case, a total of 17 component measures were originally proposed, only 13 of which were retained. Again, the initial reflected insights gathered from more than eight years of observation of how students

played the game and how the game activity functioned as an assessment tool in concrete teaching practice. Based on our experience, the selected measures all had high face validity.

Study #1

To understand the psychometric properties of the 17 original Decider measures, we began with a principal component analysis ("PCA"). We administered the Decider assessment to two different samples: the first a secondary school, which involved 379 participants (ages 13 to 14); the second a university, which involved 352 participants. Genders were evenly represented in both cases. Our goal was two-fold: (a) to examine the different components assessed by the measures, and (b) to identify potentially problematic measures for further consideration and improvement. The results of this analysis can be found in Table 1.

Table 1. PCA results for the Decider SQ assessment.

Principal Component Analysis

Sample	Measure	Component 1	Component 2	Component 3	Component 4
Secondary School	Measure 1	**0.94**	0.28	0.03	-0.10
	Measure 2	**0.91**	0.21	0.06	0.29
	Measure 3	**0.89**	0.44	0.04	-0.05
	Measure 4	**0.82**	0.11	0.50	-0.09
	Measure 5	0.20	**0.89**	0.11	-0.01
	Measure 6	0.14	**0.92**	0.16	-0.02
	Measure 7	0.59	**0.77**	0.08	-0.01
	Measure 8	0.40	**0.60**	0.08	0.16
	Measure 9	0.01	0.36	**0.91**	-0.01
	Measure 10	-0.08	0.17	**0.87**	0.40
	Measure 11	0.34	-0.22	**0.73**	-0.07
	Measure 12	-0.11	-0.19	0.08	**0.95**
	Measure 13	0.28	0.46	0.11	**0.77**
University	Measure 1	**0.98**	0.22	0.19	-0.11
	Measure 2	**0.89**	0.28	-0.02	0.10
	Measure 3	**0.72**	0.42	0.11	-0.09
	Measure 4	**0.71**	0.20	0.58	-0.27
	Measure 5	0.20	0.81	0.09	0.07
	Measure 6	0.20	**0.78**	0.18	-0.11
	Measure 7	0.70	**0.63**	-0.04	0.00
	Measure 8	0.55	**0.65**	-0.03	0.11
	Measure 9	0.02	0.34	**0.92**	0.02
	Measure 10	-0.23	0.15	**0.94**	0.34
	Measure 11	0.27	-0.08	**0.78**	-0.02
	Measure 12	-0.22	-0.16	0.01	**0.78**
	Measure 13	0.52	0.58	0.13	**0.94**

Among the 17 initial measures, PCA results lead to 13 measures pertaining to four components being selected for both samples. The percent of variance explained by those items was 83% for the Secondary School sample and 85% for the University sample. The 13 measures finally selected (see Fig. 2 above) were used to build our final three SQ components of *Jump in and know where you are*, *Know what you can do*, and *Do it and see if it works*.

Consistent with what we had conceptualized in the design stage, there appeared to be four meta-measures that summarized what is being captured by the individual measures. The measures loaded in each component also overlap greatly with our conceptualizations. The big difference between theory and data is, however, that three measures did not enter into any of the four components. Upon close inspection we found that one was correlated too highly with one of the 13 included measures; the second simply had low loadings across board; and the third did not produce enough variation among the subjects. And, to anticipate somewhat, we ultimately decided to remove the fourth meta-measure.

Among the 17 initial measures, PCA results lead to 13 measures pertaining to four components being selected for both samples. The percent of variance explained by those items was 83% for the Secondary School sample and 85% for the University sample. The 13 measures finally selected (see Figure 2 above) were used to build our final three SQ components of *Jump in and know where you are*, *Know what you can do*, and *Do it and see if it works*.

Consistent with what we had conceptualized in the design stage, there appeared to be four meta-measures that summarized what is being captured by the individual measures. The measures loaded in each component also overlap greatly with our conceptualizations. The major difference between theory and data, however, was that three of the measures did not enter into any of the four components. Upon close inspection we found that one was correlated too highly with one of the 13 included measures; the second simply had low loadings across board; and the third did not produce enough variation among the subjects. And, to anticipate somewhat, we ultimately decided to remove the fourth meta-measure.

With the four meta-components and the measures in each determined, we calculated the Cronbach alphas for all components in each sample. (See Table 2 below.) In general, the numbers are quite high and within an acceptable range, thus supporting the internal consistency of the assessment.[5] Moreover, while it would be reasonable to claim that

5 Typically, a Cronbach alpha of .70 or higher is considered acceptable in social science research. There are, however, many cases where lower Cronbach alphas are deemed appropriate. See the useful survey article: Keith S. Taber (2018), "The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education," *Research in Science Education* 48, 1273–1296. In our development of

Component 3 (in Table 2) is acceptable with respect to internal consistency, upon further reflection the decision was made to excise it from the final version of the Decider.

Table 2. Internal consistency of Decider components (Cronbach **α**).

Cronbach α for each game measure component

	Secondary School	University
Component 1	0.93	0.93
Component 2	0.90	0.88
Component 3	0.67	0.72
Component 4	0.71	0.82

In addition to those who would use the MirMe assessment system within an academic setting, it is good to remember that results from the Decider are used within corporate environments. To this end, it is useful to compare MirMe with other psychometric instruments that are likely to be found in a business setting. Probably the two most widely-used assessments employed by HR departments, recruiters, and corporate trainers alike are the Myers-Briggs Type Indicator ("MBTI") and the DiSC Model assessment ("DiSC"). Both of these come in multiple versions with varying psychometric properties, but for a rough comparison, the dimensions of the MBTI have been shown to have Cronbach alphas that range .64 and .84,[6] while DiSC Cronbach alphas range from .70 to .92.[7]

To see how the Decider correlates with other academic assessments and aptitude tests, we also collected four different measures on the two participant samples. These are summarised in Table 3 below.

---

the 15-question survey for MirMe, we achieved lower Cronbach alphas but were willing to work with the instrument in light of the nature of the task and its good test-retest validity.

6 R.M. Capraro and M.M. Capraro (2002), "Myers-Briggs Type Indicator Score Reliability Across Studies: a Meta-Analytic Reliability Generalization Study," *Educational and Psychological Measurement* 62, 590-602, p. 594.

7 As reported in: Inscape Publishing (2005), "DiSC Validation Research Report." https://www.onlinediscprofile.com/wp-content/uploads/disc-research.pdf (accessed 31 October 2016). See also the research report by Wiley on their version of DiSC, which is available at https://www.onlinediscprofile.com/wp-content/uploads/Everything-DiSC-Research-Report.pdf (accessed 2 August 2020).

Table 3. Decider correlations with other measures for sample groups.

Sample		Effective Sample Size	Validation Measure	Achieved R^2 Using MirMe Measures
Secondary School		248	PSLE score	0.184
University	University wide	68	Current GPA	0.335
			A-level average	0.248
	Analytical Skills classes	107	Final grade	0.128

We begin our consideration with the secondary school sample. For this group, we asked that students complete at least 5 trials of the game. Unsurprisingly, not all students complied, so the final effective sample size is only a subset of all students. The validation measure we collected for the secondary students was their PSLE score, taken at the end of the academic year immediately prior to entry into secondary school.[8]

For the university sample, there are two sub-samples, each of which includes students from all years and from diverse majors in the university. The first university sub-sample consists of a generic university-wide pool of students who were taking different courses at the time. There are two validation measures for this sample: the students' current GPA (i.e., grade point average, measured on a 4-point scale) and their A-level scores (taken before entering university). The second university sub-sample consists of students currently taking the Analytical Skills course ("AS-course"), a course in critical thinking required for all students at the university. The AS-course is typically taken during the first year of university (92 participants, or 86% of the sub-sample), but due to scheduling factors students from all four years are represented (the remaining 15 participants, or 14% of the sub-sample). There is one validation measure for this sample: the student's final grade in the AS-course. We asked all students to complete at least 3 trials of the game; however, not all complied, which resulted in a reduced effective sample size.

We used mainly regression analysis to analyse the predictive validity of MirMe Decider measures. In each sample or sub-sample, we first conducted a Generalized Linear

8 The PSLE, or "Primary School Leaving Examination," is a mandatory national examination that all students in Singapore take at the end of Primary 6. English, Mathematics, Science, and Mother Tongue are assessed, with approximately 2 hours dedicated to each subject. The PSLE determines whether a Singapore student is permitted to move on to secondary school, determines which secondary school the student may enter, and determines the academic stream within which the student is placed. To a significant degree, who gets to attend university is determined for Singaporeans at the age 12.

Model ("GLM") analysis. The goal was to select the best model, measured by Bayesean Information Criterion ("BIC"), based on the different measures of the Decider. We then calculated the adjusted r^2 value of this model. The results are shown in the last column of the table. As is usually the case with psychometric instruments, the Decider could possibly achieve different levels of predictive validity depending on the sample and validation measure. The results suggest that the Decider tracks student academic performance, with a slightly greater explanatory power for university-age students than secondary school students.

Study #2

We then wished to consider how MirMe performed relative to other intuitively important measures within student life. The Decider was administered to 128 university students, 45 male and 83 female. Data collected included A-level results, academic GPA to-date, and the co-curricular activities ("CCAs") both prior to and subsequent to entering university. The understanding of what constitutes a CCA was broad, including clubs, charitable organisations, professional organisations, and in general any student-centric, organised activity that falls outside the standard academic courses. The CCAs were coded along two dimensions: (1) *group* versus *individual*, and (2) *cognitive* versus *non-cognitive*. Thus, chess club was coded as (*individual* + *cognitive*), whereas dragon boating was coded as (*group* + *non-cognitive*). The results are summarised in Table 4.

Table 4. MirMe correlated with A-level, GPA, and CCA for university students.

Measure	Correlated with...	R^2/Adjusted R^2, pvalue
MirMe	A-levels	$R^2 = .211$, $p < .001$
MirMe	GPA	$R^2 = .265$, $p < .001$
A-levels	GPA	$R^2 = .209$, $p < .001$
MirMe + A-levels	GPA	$R^2 = .361$, $p < .001$
MirMe	Pre-U CCA – cognitive	$R^2 = .238$, $p < .001$
MirMe	Uni-CCA – cognitive	$R^2 = .391$, $p < .001$
MirMe	Pre-U CCA – group	$R^2 = .178$, $p < .01$
MirMe	Uni-CCA – group	$R^2 = .182$, $p = .039$

The results of suggest that the Decider may be a more accurate predictor of university academic performance (as measured by GPA) than A-level results. Inspection

of Table 4 further reveals that the Decider, when taken in conjunction with A-level results, yields an even better predictor of student GPA than either measure taken individually. This is consistent with the hypothesis that whereas MirMe focuses upon the assessment of skills, the A-levels place greater emphasis upon content knowledge, both of which are important contributors to GPA.

Furthermore, according to this study, the Decider is also capable of predicting, both retrodictively and prospectively, the CCAs that participants engage in. We believe that this may be of interest to university administrators who wish to encourage and support diversity in student interests and thereby enhance the students' educational experience. It is further hypothesized that such information may be relevant for highly-desirable academic and business outcomes such as team optimisation.

Study #3

As a follow-up to Study #2, a brief investigation was conducted to discover how the Decider correlates with other well-known psychometric assessments. The results of this Study #3 have been published.[9]

In the study, participants were secondary school students who took both the Decider and the TIPI ("Ten Item Personality Inventory") as well as academic data such as the PSLE. The TIPI is a short pencil-and-paper test using the 5-Factor ("Big 5") dimensions commonly used in psychology research.[10]

In brief, the results of this study suggest that the relationship between the Decider and TIPI is weak and that the two measures are probably orthogonal to one another. This is in line with expectations given the focus of the Decider SQ assessment.

Cultural Theory survey and experimental results

In addition to the abstract game at the heart of MirMe SQ assessment, the system also includes a questionnaire comprised of 15 yes/no questions. This survey is intended to reveal the participant's 'cultural profile', which can be thought of as a snapshot of how a

9 Yong, M.S.K., and Shin, Y.H. (2015), "Psychometric Assessment of 21st Century Employability Skills: Situational Intelligence and Social Factors." *Humanities and Social Sciences Research Programme (HSSRP),* 189–196.

10 For details, see Sam Gosling's website: https://gosling.psy.utexas.edu/scales-weve-developed/ten-item-personality-measure-tipi/ (accessed 29 June 2020). For the original TIPI paper see: S. Gosling, P. Rentfrow & W. Swann, Jr. (2003), "A Very Brief Measure of the Big Five Personality Domains," *Journal of Research in Personality* 37, 504-528.

participant prefers to arrange his or her social relations. The survey draws upon Cultural Theory ("CT"), a sociological theory introduced by Dame Mary Douglas.[11]

According to CT, there are four basic ways of organising social relations: hierarchy, egalitarianism, individualism and fatalism. These four ways of organising are described by assigning 'high' and 'low' values to two dimensions of social life: the extent to which people are incorporated into a larger social setting ('Group') and the degree to which people are regulated and ranked ('Grid'). Hierarchy combines high stratification (+Grid) with a high degree of collectivity (+Group); individualism scores low values for both stratification (-Grid) and collectivity (-Group); egalitarianism exhibits high collectivity (+Group) but low stratification (-Grid); and fatalism is characterised by high stratification (+Grid) and low collectivity (-Group).

Each of the four ways of organising come packaged with a distinct pattern of perceiving, justifying, reasoning, acting, and feeling.[12] Taken together, each pattern constitutes a 'way of life'. The various predispositions (of beliefs, values, perceptions, etc.) endogenous to a way of life may be called its 'cultural bias'.

For some intuitive examples of the four ways of life, we can say that members of the Singapore civil service score high on both grid and group and are thus hierarchists. Members of several utopian communities (e.g., the Twin Oaks community in Virginia or an Israeli *kibbutz*) are egalitarians. Self-made entrepreneurs and capitalists Bill Gates and Warren Buffett are individualists. And lastly, non-unionized graduate students, whose lives are subject to the capricious whims and dictates of their professors, are typical fatalists.

As there was no CT survey hat would be either suitable or available for incorporation into MirMe, it was necessary to develop and test various question items and determine appropriate internal cutoff values.

---

11 For an accessible introduction to CT, see Mary Douglas, Michael Thompson, and Marco Verweij, "Is time running out? The case of global warming," *Daedalus* 132, 2 (2003): 98–107. Key theoretical publications include: Mary Douglas, "Cultural Bias," *Occasional Paper No. 35* (London: Royal Anthropological Institute, 1978); Mary Douglas (ed.), *Essays in the Sociology of Perception* (London: Routledge, 1982); Michael Thompson, Richard Ellis and Aaron Wildavsky, *Cultural Theory* (Boulder, CO: Westview Press, 1990); and Michiel Schwarz and Michael Thompson, *Divided We Stand: Redefining Politics, Technology and Social Choice* (Philadelphia; PE: University of Pennsylvania Press, 1990).

12 The roots of the CT ways of life go deep. Even in early childhood we find children appealing to forms of ethical reasoning characteristic of distinct ways of life. See: Mark Nowacki (2011), "Social virtues within and across cultures: Against the idea of universal rationality," *TRANS: Proceedings of Knowledge, Creativity and Transformations of Societies* 17, 8–20.

Method

Participants. Three study groups were formed: Group A, with 123 primary school students (ages around 11); Group B, with 135 secondary school students (ages around 13); and Group C, with 125 university students (ages average around 21).[13] The participants were drawn from Singapore schools and was evenly divided between males and females.

Procedure. The study was comprised of two parts: a survey and an experiment. In the survey, participants were asked to fill in a 15-item questionnaire, in which six items were meant to measure their standings in the Grid dimension and nine in the Group dimension. Because there was no pre-existing Grid scale suitable for school children, we developed the items ourselves (e.g., "I believe that whether a person will be successful in life has a lot to do with what type of family the person is from" and "when taking the stairs at school, I almost always use the correct side of the staircase"). The highest possible total score in the Grid scale is 11 and the lowest is 5. Items in the Group scale were adopted from a scale developed by Singelis (1994) (e.g., "I believe that what my friends want is more important than what I want" and "I am comfortable with being praised or rewarded in front of my friends").[14] The highest possible total score in the Group scale is 15 and the lowest is 6.

Approximately two weeks after the survey was conducted, an experiment was carried out with the following procedure: First, a teacher/experimenter asked students in a class—class size ranged from 16 to 19—to guess a number X between 1 and 20. In all sessions of the experiment, X was set at 19. After students wrote down their guesses, the experimenter revealed X. Instead of the typical rule that the closer one's guess is to the target, the better one's performance, the reverse was applied. Thus, guessing "1" would actually result in the best performance and "19" the worst. After announcing the rule, the experimenter asked the two best performers over and presented them with rewards: a box of chocolate with 20 individually wrapped chocolate balls. The experimenter then withdrew one ball from the box, presenting it to the second best performer and leaving the rest to the best performer. In cases of ties, who would get the larger reward or any reward at all was decided by the experimenter

---

13 The MirMe test has been used within other academic studies, notably: Poh Sun Seow, Gary Pan & S. Grace Koh (2018), "Examining an experiential learning approach to prepare students for the volatile, uncertain, complex and ambiguous (VUCA) work environment," *The International Journal of Management Education* 17(1), 62-76. Another (2021) study involving 109 working adults was undertaken for the Singapore Ministry of Manpower and will be further expanded in the near future.

14 T. M. Singelis (1994), "The measurement of independent and interdependent self-construals," *Personality and Social Psychology Bulletin* 20, 580-591.

with an arbitrary rule, such as who was taller or shorter. Finally, after the reward presentation, all students were asked to rank-order the fairness of the following five reward distribution options:

Hierarchist (HIER)	Everyone gets at least some, then people who did better get more.
Egalitarian (EG)	Everyone gets the same.
Individualist (IND)	People who did better get more candy.
Fatalist 1 (FT1)	The way how the instructor did it.
Fatalist 2 (FT2)	Any method is fine, even no candy for anyone, so long as nobody gets more than I do.

Participants were instructed to give "1" to the fairest option and "5" to the least fair one. We developed the options in such a way that each would, in our opinion, be identified most by a certain CT-type. There were two fatalist options, because we consider both viable for a fatalist. In addition to those five options, students were also asked to indicate if their favorite choice was not in the list. Seven participants did so but they did not elaborate on what those alternatives were. Overall, we consider that the five options offered a fairly comprehensive coverage of the possible distribution options in the experiment.[15]

Reliability. A test-retest exercise was conducted with Group A, who were presented with the same survey questions three to four weeks after the first administration of the questionnaire.

Results

Survey. The distributions of participants' scores in the Grid and Group scales can be seen in Tables 5(A,B,C). All three groups exhibit a slight but statistically significant correlation between the two dimensions: Group A, $r = 0.266$; Group B, $r = 0.213$; Group C, $r = 0.136$; $p \leq 0.015$ for all three groups. As age increases, the two dimensions becomes less and less correlated. Assuming that we are measuring the right underlying variables (or constructs), Grid and Group tend to be perceived as more separate and distinctive when people get older. There has been little or no discussion in the CT literature about this

15 While the ranking experiment was also run with Group C, only the results from Group B are discussed in this paper. Data from Group C will appear in a forthcoming publication. In brief, the results from Group C are similar to that of Group B, in that CT way of life appears to correlate with the participant's preferred distribution method.

phenomenon. This opens up a number of future research questions. Are they supposed to be independent or correlated? If the latter, is the correlation positive or negative; and how will it change across different age and culture groups? The survey results are suggestive, but it is likely that different approaches will be needed.[16]

Table 5(A). Distributions of Participants' Scores in Grid & Group Scales – Group A (Test 1).

Grid Scale	Group Scale										
	Score	6	7	8	9	10	11	12	13	14	15
Score	Frequency	0	2	10	14	29	21	23	14	6	4
11	0										
10	9				1	2		4		2	
9	25			1	3	5	5	8	1	1	1
8	25			2	3	4	3	4	4	2	3
7	34		1	3	2	9	7	4	7	1	
6	27		1	3	5	7	6	3	2		
5	3			1		2					

Table 5(B). Distributions of Participants' Scores in Grid & Group Scales – Group B.

Grid Scale	Group Scale										
	Score	6	7	8	9	10	11	12	13	14	15
Score	Frequency	0	0	7	19	27	27	29	17	7	2
11	0										
10	12				2	1	1	4	1	3	
9	20			1	1	4	3	6	5		
8	44			1	4	12	10	10	6		1
7	39			3	8	4	10	6	4	4	
6	17			1	4	6	2	2	1		1
5	3			1			1	1			

16 Survey-based instruments for scoring Grid and Group, while practically unavoidable in many situations, have a number of important limitations. Structured observation particularly holds promise as an alternative method for testing CT. See: Marco Verweij, Marieke Van Egmond, Ulrich Kühnen, Shenghua Luan, Steven Ney & M. Aenne Schoop (2014), "I disagree, therefore I am: how to test and strengthen cultural versatility," *Innovation: The European Journal of Social Science Research*, 27(2), 83-98. This article builds on suggestions in Marco Verweij, Shenghua Luan & Mark Nowacki (2011), "How to Test Cultural Theory: Suggestions for Future Research," *PS: Political Science & Politics*, 44(4), 745-748. A representative example of such an experimental approach is reported in: M. Aenne Schoop, Marco Verweij, Ulrich Kühnen, & Shenghua Luan (2020), "Political disagreement in the classroom: testing cultural theory through structured observation ,"*Quality & Quantity* 54, 623–643.

Table 5(C). Distributions of Participants' Scores in Grid & Group Scales – Group C.

Grid Scale	Group Scale										
	Score	6	7	8	9	10	11	12	13	14	15
Score	Frequency	3	7	13	18	21	29	14	12	7	1
11	2						1	1			
10	17			2	3	3	2	3	3	1	
9	34		2	6	3	7	9	1	3	3	
8	37	1	2	3	7	6	8	5	3	1	1
7	28	1	2		5	4	7	4	3	2	
6	6	1		2		1	2				
5	1		1								

We distinguish "high" and "low" scores in each dimension using the median cutoff. The number of participants classified for each CT-type: hierarchist, egalitarian, individualist, and fatalist, and their scores can be seen in the four shaded areas in Tables 1(A,B,C), starting from the upper right corner and clockwise, respectively. Median cutoffs, of course, are just one way of making sense of participants' scores. Different cutoffs might result in different classification results, and all such classifications should be understood only relatively (i.e., a hierarchist is a person who scored relatively high in the Grid and Group dimensions compared to others in the same sample).

To test the effect of changing the scores used to define "high" and "low" scores in the Grid and Group dimensions, three types of cutoff values were assayed. "Liberal" means that the cutoff values are relatively low; as a result, a higher proportion of people are classified as belonging to a certain CT type; "conservative" means that the cutoff values are relatively high; and "median" means that they are sort of in the middle. As may be observed in the three charts produced in Figures 3(A,B,C), choice of cutoff values does not seem to affect the results much, supporting the general robustness of our results.

Figure 3(A). Results of applying median cutoff values for Grid and Group.

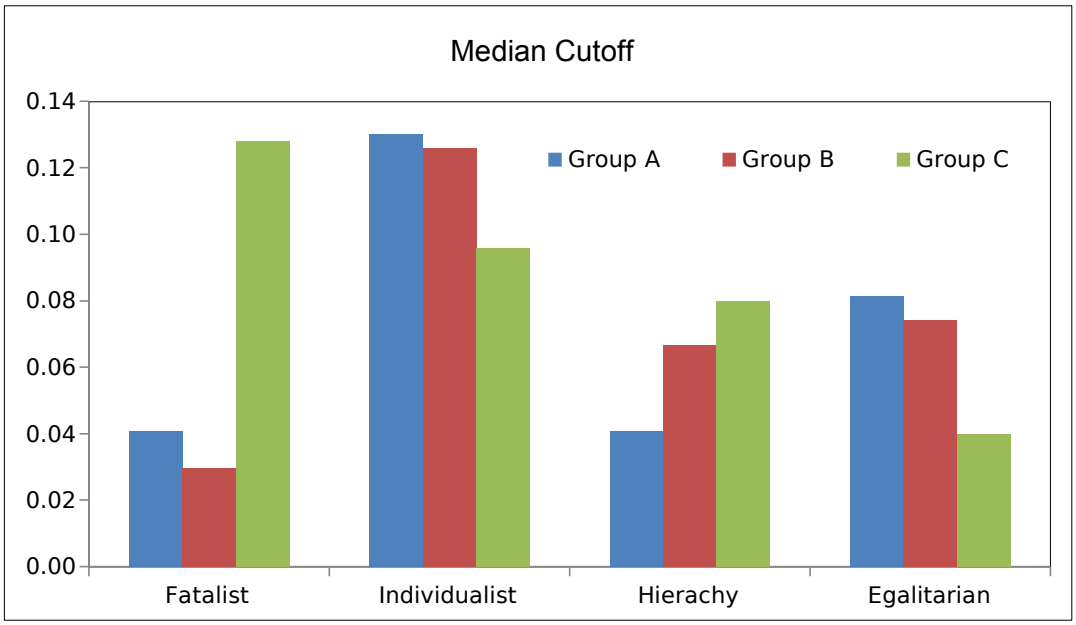


Figure 3(B). Results of applying liberal cutoff values for Grid and Group.

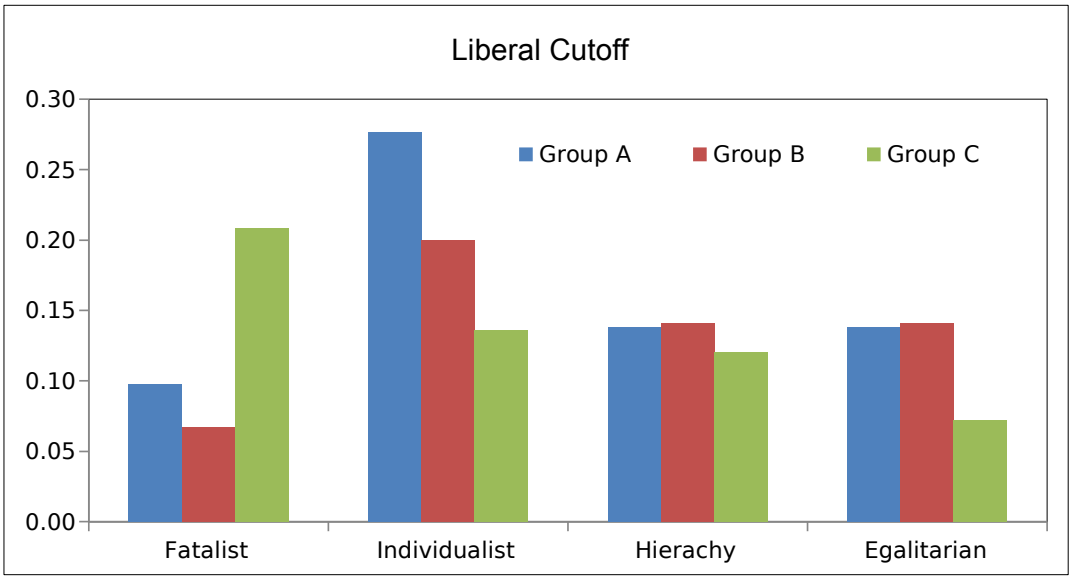
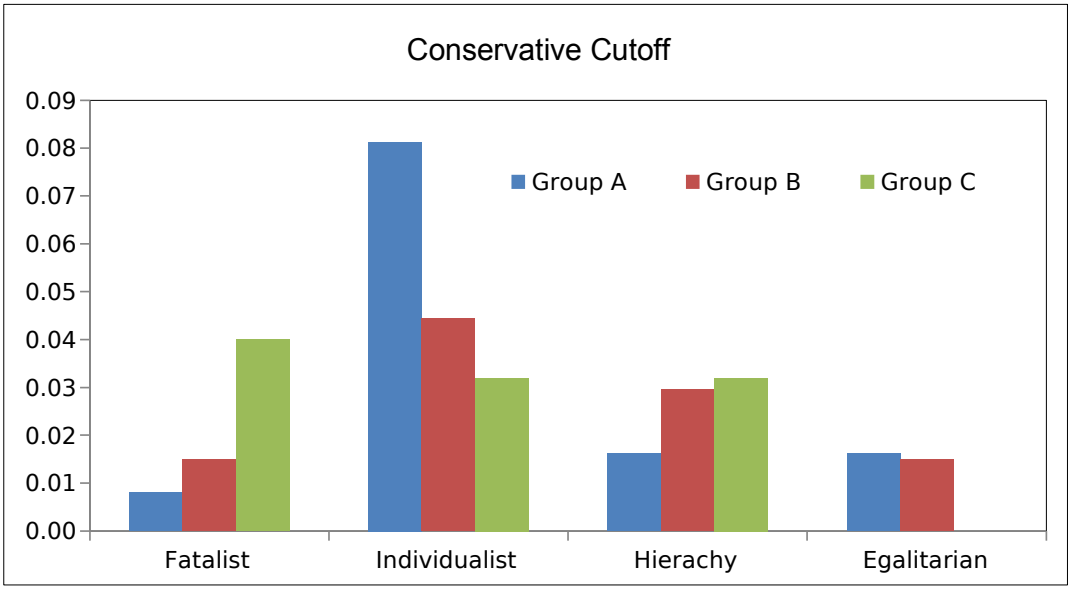


Figure 3(C). Results of applying conservative cutoff values for Grid and Group.



Inspection of Figures 3(A,B,C) reveals that: a) there are proportionately more fatalists and hierarchists in Group C compared to Group A and Group B; b) there are proportionately fewer individualists and egalitarians in Group C compared to Group A and Group B; c) the two younger samples, that is, Group A and Group B, resemble each other more than they do Group C. Apparently, stepping into adulthood can change one's internal views about the world.

The differences in proportional representation of CT types found in Group C versus the two younger samples may be driven primarily by the higher score (on average) of Group C in Grid, as shown in Table 6. Meanwhile, Group C also tends to score lower in Group than the two younger samples. While maturation may be the best explanation for the proportional shift, it is possible that the results are subject to selection bias, as not all primary and secondary school students end up attending university. Without further longitudinal data (which will be difficult to secure), it is impossible to rule out the possibility of some selection bias.

Table 6. Mean and standard deviation of Grid and Group scores.

Sample	Mean Score		Standard Deviation	
	Grid	Group	Grid	Group
Group A	7.561	10.919	1.300	1.818
Group B	7.719	11.044	1.195	1.634
Group C	8.248	10.392	1.182	1.991

Experiment. The rank order of each distribution option, averaged across all participants classified under a certain CT-type, can be seen in Table 7. The rank of the fatalist (FT) option was based on the average rank of the two fatalist options (FT1 and FT2). This approach is justified by the relatively small difference between the average ranks of the two options (the average rank for FT1 and FT2 was 3.92 and 4.49, respectively, over all participants).

Table 7. Average Rank Order of Each Distribution Option under Different CT-types.

CT-type	# of Participants	Average Ranking Order				Rank Order by Statistical Test Results*
		HIER	EG	IND	FT	
Hierarchist	19	1.95	2.47	2.58	4.00	HIER>EG=IND>FT
Egalitarian	19	1.74	2.05	2.89	4.16	HIER=EG>IND>FT
Individualist	27	1.81	2.52	2.41	4.13	HIER>IND=EG>FT
Fatalist	9	1.11	2.56	2.33	4.50	HIER>IND=EG>FT
Unclassified	61	1.57	2.52	2.36	4.27	HIER>IND=EG>FT

*Results based on non-parametric permutation tests with 50,000 permutations; $p = 0.05$ was applied.

We see from the table that in general the hierarchist (HIER) option was ranked the highest and the FT option the lowest, with the egalitarian (EG) and the individualist (IND) options in the middle. However, differences in rank orders did exist among different CT-types. Specifically, for egalitarians, the rank difference between the HIER and EG options was not statistically significant, but the difference between the EG and IND options was. Thus, the EG option was clearly preferred by the egalitarians to the IND option. The reverse tendency was observed with individualists, although it was not statistically significant. Fatalists share the same preference structure as individualists, with more pronounced liking

for the HIER option and disliking for the FT option. Given the small number of fatalists in our sample (9), however, the reliability of these results is questionable. Finally, the result that the HIER option was the favorite not only for the hierarchists but for all participants came as a surprise to us. A possible explanation is that the HIER option was seen as an *ex post facto* compromise strategy that would mitigate the risk of receiving no reward. According to the experimental design, only the top two participants received any reward, and the top participant received 19 chocolates to the runner-up's single chocolate. HIER would guarantee a share for the (jealous?) majority of participants who received none of the chocolate.

Overall, correspondences are found between the survey results (i.e., classifications of the four CT-types based on scale scores) and the experiment results (i.e., rank orders of the reward-distribution options). They not only support the notion that individual differences in CT-type can be used to explain why people's opinions can sometimes differ drastically, but also show that the understanding of CT may benefit from further testing employing multiple methods.

Reliability. Two measures of reliability were used: a) correlation between the test and retest scores; and b) agreement between the scores. Agreement is understood as follows: if the participant answered "yes" on an item at Time 1, what is the probability of the participant saying "yes" again at Time 2. The numbers are averaged across all selected items and all subjects. Since we have only binary data in each survey item, the correlation is somewhat deflated. The agreement numbers are good. See Table 8.

Table 8. Test-retest correlation and agreement.

	Correlation	Agreement
Grid	0.495	0.764
Group	0.626	0.732

Distribution of scores and CT types. Table 9 records the distribution of scores for the re-test administered to Group A. Table 9 should be compared with Table 5(A) above, which shows the results from the first test. Inspection of the two tables reveals that, though some changes have occurred, the overall distributions remain largely intact. The same is true of the proportions of the different CT types, with the exception of the fatalists. Upon further review, the fatalists appear to be the least reliable participants in the surveys. (For instance,

fatalists tend to adhere more weakly to their distribution options in the experiment described above.) Moreover, while there is some movement between test and re-test, with the application of the median cutoff standard discussed earlier, none of the subjects crossed over from one CT type to another. This further testifies to the overall reliability of the subjects' survey scores.

Table 9. Distributions of Participants' Scores in Grid and Group Scales – Group A (Test 2).

Grid Scale	Group Scale										
	Score	6	7	8	9	10	11	12	13	14	15
Score	Frequency	0	3	8	9	21	31	21	19	7	4
11	0										
10	7					1	1	3	1		1
9	20				2	2	8	5	1	1	1
8	38			3	1	8	8	5	7	4	2
7	36			4	1	7	8	7	7	2	
6	20		2	1	4	3	6	1	3		
5	2		1		1						

After winnowing down the initial list of 33 survey questions, we were left with six Grid questions (Cronbach alpha ≈ 0.3) and nine Group questions (Cronbach alpha ≈ 0.48). These are the 15 survey questions used by MirMe.

Remarks on the MirMe SQ assessment system

The research results support the claim that the LogicMills MirMe Decider is an effective assessment for Situational Intelligence. As a psychometric tool, MirMe has a number of unique advantages. First, MirMe is the only comprehensive testing instrument for 21st century skills. Second, MirMe assessments are impossible to game, precisely because they use games. Unlike traditional tests like MBTI and DiSC, where re-testing leads to the participant being able to determine the desired outcome, the Decider presents a consistently-assessed yet changing environment. It is impossible to memorize a good way through the Decider, in much the same way that it is impossible to memorize all the best moves in chess. Third, the Decider is quick to administer: participants can navigate the assessment in under 45 minutes. Fourth, because the Decider is based on a game, it is non-threatening and likely to elicit a meaningful and engaged response.

Future research

While the results obtained thus far are positive, we suggest that future research may may focus on the Decider's performance relative to the SAT ("Scholastic Aptitude Test"), which is widely used by North American universities. As a preliminary consideration, we note that published studies of the SAT typically report that the percent of variance in first-year GPA predicted for by the SAT (Reading + Mathematics) score is between 13% and 20%. Using the College Board's own study of the latest version of their test, the correlation between SAT and first year GPA has adjusted $r = .51$, entailing a percent variance of 26% (i.e., adjusted $r^2 = .26$).[17] In line with the results of Study #2, the Decider is predicted to perform as well or better than the SAT, for the Decider accounts for 26.5% (adjusted $r^2 = .265$). Furthermore, as appears to be the case with the A-levels, we hypothesize that the explanatory power of the Decider plus the SAT will be greater than either measure alone.

Authors: Shenghua Luan
Center for Adaptive Behavior and Cognition
Max Planck institute for Human Development
Berlin, Germany

Mark R. Nowacki
School of Social Science
Singapore Management University
Singapore

Contact: mark@logicmills.com

17 See Westrick et al. (2019), "Validity of the SAT® for Predicting First-Year Grades and Retention to the Second Year," *CollegeBoard,* https://collegereadiness.collegeboard.org/pdf/national-sat-validity-study.pdf (accessed 14 July 2020).