Executive Summary-Recommendations

This document is the result of the Admissions Committee members' discussion and evaluation of DEMRE´s proposal (2018). The Active Committee members were Salomé Martínez, Alejandra Mizala, Verónica Santelices and Rebecca Zwick[1]. In the following pages, each one of DEMRE´s proposals is discussed independently. Regardless of the decision reached on each of the specific changes proposed, it is important that these changes to the PSU battery are made in a progressive manner, with modifications introduced sequentially and, ideally, one at a time. Furthermore, proper and advanced information should be provided to secondary education teachers and students, families and higher education institutions.

Committee members agreed that it is a priority to address the weaknesses of the Math test and Language test. It is pressing to split the Math test into two separate tests of differentiated content (basic and advanced) and to exclude current proxies of writing (connectors and the writing plan) from the Language test. In parallel, the implementation of a Writing test should be considered in order to signal schools the importance of teaching writing during secondary education. While all four members agreed on the need to have a mandatory basic math test, there was no agreement on whether the advanced Math test should be mandatory for all students or just for students applying to all math-intensive programs.

The Committee members did not come to an agreement on whether Item Response Theory should be implemented to analyze and score the PSU. Advantages and disadvantages were weighed differently by different members of the committee. They agreed, however, on the need to study its implementation and test different models, estimation procedures and potential communication challenges that transitioning into such a system would pose to DEMRE staff and other relevant stakeholders. It should be noted that under the IRT models that would be useful for the PSU, scores would no longer have a simple relationship to the number of correct responses.

[1]The appendix to this document details the meetings and communications held by committee members between May and December of 2018, and between committee members and SUA staff during the same period of time.

The benefits of replacing the current Science test with three separate Science tests, each measuring Biology, Physics and Chemistry, are less clear and require more study before this idea is implemented. Some committee members supported DEMRE´s proposal to report the discipline of advanced content chosen by the student and the score obtained by the examinee on the items of that specific discipline (combining basic and advanced items) in addition to the test total score. Other committee members, though, suggested increasing the number of items in the common section of the test in order to reach a reliability level that would allow reporting the basic content section score separately and independently from the score on the advanced section.

Salomé Martínez
Alejandra Mizala
María Verónica Santelices
Rebecca Zwick

Proposal 1. To review and clearly define intended uses and interpretations of the PSU battery.

In order to address the first point raised by DEMRE´s document (April 2018), we suggest that SUA lead a study exploring the Theory of Action, or Logic Model, of the PSU scores, including the intended and unintended uses of the PSU scores expected by different stakeholders including test designers and test score users. The importance of defining the use of test scores is raised by the Standards for Educational and Psychological Testing (American Educational Research Association [AERA], American Psychological Association [APA] & National Council on Measurement in Education [NCME], 2014) and this definition should explicitly be undertaken for the Chilean university admissions test.

The study may be organized by SUA professionals, in collaboration with staff from CRUCH, or analysts from DEMRE, but should be informed by the account of leaders from the different stakeholder groups, including higher education authorities, decision makers from the Ministry of Education and, perhaps, members of the secondary education sector (high schools).

In order to define the "expected uses" of an evaluation system, the measurement or test may be considered as the equivalent of an educational program or intervention. This concept or device, characteristic of the literature of the Evaluation of Programs and Policies (Rossi, Freeman & Lipsey, 2003), explores in depth the expectations of those who participate in the design of the evaluation and of those who actively use its scores, as well as the mechanisms through which the goals, these being the expected uses, are expected to be achieved. The important consequences of assessments such as the Chilean PSU for individuals suggest that the Theory of Action should be explicitly stated and examined.

Defining the uses that should be given to the information obtained from an instrument or measure is essential in order to educate potential users in relation to the appropriate interpretations of the information that the instrument delivers (NCME Statement, 2018). The Theory of Action, or Logic Model, also serves as the basis for any validation effort. An adequate validation process requires that the uses of the

instrument scores be defined and made explicit as early as possible. The concept of validity, understood in the light of the Standards (AERA, APA & NCME, 2014) and Kane (2016), presents a validation process “as a complex argument” using quantitative and qualitative evidence, but always conceived and implemented for a specific "use" and "interpretation" of the scores. From this perspective, it is not possible to validate an instrument, or test, in a vacuum, ignoring the uses that are given to its scores in the psychological or educational world. In addition, the explicit definition of this "use" and its active communication will allow for a better definition of the criteria to be considered in studies of predictive validity and validation of standards.

The Theory of Action is "the process by which the components of the program are supposed to affect the results and conditions in which these processes are believed to operate" (Donaldson, 2007, p.22). When the Theory of Action is not available, experts advise using multiple sources to develop it (Rogers, Petrosino, Huebner & Hacsi, 2000). These sources include the implicit theories of those who were close to the design and are close to the implementation of the program, documentation of program activities and observation and exploration of key assumptions. The objective is to create a simple, but complete program theory, or a plan, that relates the activities (or components) of the program, or assessment, with the expected objectives. This Theory of Action can be used to prioritize evaluation questions (Donaldson & Gooler, 2003). Most program theories are summarized in a diagram that represents a causal chain (Rogers et al., 2000). Once the Theory of Action or Logic Model is established, the attention should be focused on exploring the empirical evidence that supports (or not) the validity argument (s) implicit in that theory. The validity arguments will be given by the "uses" or "expected effects" specified in the theory of action.

Proposal 2. To review the scope of content that is considered in the PSU test battery.

Regarding DEMRE´s proposal to review the scope of content considered in the PSU test battery, we believe the issue should be addressed using an in-depth perspective. We further discuss this issue in the next section for Mathematics (section three) and we address some related issues for Language and Science in sections four and five. The definition of the content to be covered by the test should be examined together with relevant stakeholders, such as higher education authorities, representatives from high schools and the Ministry of Education, and should not necessarily be limited to the contents that are currently included in Chile´s Secondary Education National Curriculum.

The content of the test should also correspond closely to the intended uses and expected consequences defined by the study suggested above; that is, it must be consistent with the intended uses. For example, if the agreement is to emphasize the predictive use of the test scores, aiming to predict academic success in universities, then stakeholders should decide what content is best suited to predict university performance. If the goal is to provide diagnostic information about certain general abilities to all types of higher education institutions, stakeholders from universities and technical institutions should agree on those general abilities and also on the level of achievement expected either for admissions, remediation and/or any other type of educational decision. We describe some of the studies conducted during the last ten to fifteen years on these two potential uses here below.

Predictive power of the current PSU test battery, as reported by Manzi et al. (2010) seems somewhat below that of similar tests in the United States (Mattern, Patterson, & Kobrin, 2012) and Sweden (Wikström & Wikström, 2017)[2]. However, there is not

[2]Of the two main admission tests for American universities, the SAT and the ACT have been validated by relating the performance in those tests with academic performance measured through the GPA of the first and fourth years at the university, and by graduation and persistence rates (Geiser & Santelices, 2007; Zwick, 2017). In general, the correlation between the score in the SAT (Verbal or Critical Reading and Mathematics combined) and the first year grades fluctuates between 0.3 and 0.6, depending on the characteristics of the studies, with an average of 0.4 (Ramist, Lewis &McCamley-Jenkins, 1994; Zwick 2002). Mattern, Patterson, & Kobrin (2012) found a correlation of SAT-Mathematics and first-year college Mathematics courses of 0.52. The SUA implemented a *predictive validity* study and reported a correlation corrected for

sufficient information to have a full understanding of the predictive power of the current PSU since the methods used by official analyses so far have been limited to the average correlation between test scores and university first-year grade by institution. In addition, validity studies are not conducted systematically. The last predictive validity study including all CRUCH institutions was conducted using data from 2015 (Consejo de Rectores, 2018).

Validity research should be undertaken systematically by the organization overseeing the quality of the test and/or in charge of its design and administration. Moreover, validity research should be conducted examining not only the relationship between test scores and first year GPA, but also other academic and social indicators of successful transition to higher education. These types of analyses require that researchers and professionals from SUA, or other research centers and universities, have access to the academic and social indicators of students from all institutions using the test scores. The information should be made available with no personal identification upon request from interested parties. Currently, it would be possible for the SIES (Servicio de Información de Educación Superior) to collect the relevant information fields as part of the ongoing process through which it collects enrollment and student-level information, regarding number of credits registered and number of credits approved by the students.

During the last ten years, test designers have emphasized aligning the content of the test to the Secondary school curriculum of the scientific humanistic track without giving up the predictive power of the test scores. However, the Pearson Report (2013) criticized the extent to which PSU actually covers the secondary curriculum. Another important issue to address is that the PSU content currently does not align with the vocational track curriculum, which accounts for around 45% of secondary school enrollment.

---

range restriction of 0.58 for 2013, 0.62 for 2014, and 0.56 for 2015 between the Mathematics PSU and the students' first-year cumulative weighted average (SUA, 2017). The same study found smaller correlations for the Language PSU module during the same period (0.23 in 2013, 0.20 in 2014, and 0.18 in 2015).

Proposal 3. To study the possibility of splitting the current Mathematics Test into two tests with different content and difficulty levels.

The high difficulty of the PSU mathematics test (PSU-M) is shown in the report provided by DEMRE (2018a; Section. g.), where a skewed distribution of the number of correct answers can be observed in the applications of the test from 2004 to 2018. For instance, in 2016, a score of 450 points in the PSU-M corresponds to only 19 correct answers out of 75, and the median of 496 points, corresponds to only 22 correct answers. Since each item has 5 alternatives, these scores are close to random answering (which corresponds to 15 correct answers). It should be noted that 450 points is the minimum average PSU score required for applying to a program of a SUA participating university. Due to the skewed nature of the distribution, the normalization process used to compute scores generates a situation where the students whose scores are above the median end up with small differences in the normalized scores, even though they have relevant differences in the number of correct answers. This is problematic, since this subpopulation is the one relevant for selection purposes. It should also be added that normalization is a transformation done to compare different cohorts and does not affect the ranking.

The DEMRE proposal relies on a study carried out using the 2013 cohort (DEMRE, 2017), which analyzed the distribution of the scores on the PSU-M for different subpopulations of students, also distinguishing by type of content. The authors distinguished: Basic Content (BC), corresponding to Mathematic Minimum Obligatory Contents for1st and 2nd grade of the Chilean High School Curriculum, and Advanced Contents(AC), corresponding to Mathematic Minimum Obligatory Contents for 3rd and 4th grade of the Chilean High School Curriculum. In this report, the authors study the reliability, difficulty and predictive validity of the two subsets of items. While this study addresses many important issues related to the proposal, one important shortcoming is that it was conducted using the cohort of 2013, when correction for guessing was still being implemented. Nonetheless, the conclusions of the study are still probably true for current PSU applications. Indeed, the CMM report (Lacourly, San Martín, Amaya & Uribe, 2017), where a methodology is developed to compare two different test modalities (with and without correction for guessing),

shows that in general the information obtained in both scenarios is highly correlated. Still, we strongly suggest carrying out a study with the new data.

Test difficulty.

DEMRE's *Evidencia preliminar para la reducción de contenidos en la PSU de matemáticas* (2017), shows that the test score distribution is highly asymmetrical for students belonging to the Scientific-Humanistic (SH) and Technical-Professional (TP) curricular branches of the municipal (M) and private-subsidized (PS) schools, which account for about 90% of enrollment. Besides, it shows there is an important gap in performance between students belonging to the SH and the TP branches. When only considering the BC items, the score distribution remains nonsymmetric, but as it is shown in Table 4, there is a small increase in the percentage of correct answers for the different sub-populations of students.

Regarding the students from the non-subsidized private schools (P), the PSU-M test shows anomalous behavior, since its distribution is flat. This may be related to the fact that the tasks requested by the test can be trained and that the test has low difficulty for this population. The average percentage of correct answers for this group is 65, which is much higher than the 32 percent of correct answers corresponding to students belonging to the SH branch of PS schools, which is the second highest-performing group, according to curricular branch and type of school.

Predictive value

One important factor to take into account when deciding to modify the contents of a test is whether this decreases its predictive validity. In DEMRE (2017), the authors measure academic performance by considering first-year university students' GPA. Table 7a (page 31), shows that for all students attending SUA Universities, the Math test has a correlation of 0.19 with the GPA, and of 0.18 when only the BC items are considered. In contrast, the correlation of first-year GPA and High School GPA (NEM) is 0.31. It should be noted that the correlation of the PSU-M is not consistent with the one reported in SUA (2018). In this report, the correlation of the PSU-M and

the first-year GPA is 0.26 (page 40 of the report). This discrepancy is due to methodological differences[3].

In DEMRE (2017), different regression models are estimated. It is reported that the BC and AC items do not make a significant contribution to the variance of the first-year GPA in a number of career programs. The study concludes that the predictive value of the Math test is low (for instance according to Cohen, 1988) and if only the BC items are considered, the predictive value of the test does not significantly change.

Conclusions and recommendations

The committee proposes repeating the DEMRE's 2017 study considering the 2016 PSU cohort, and so we request that the data needed be made available by SUA. In this new study we recommend individualizing career programs that may be of interest to high-performing math students, such as Civil Engineering, Math and Physics degrees. Itis also relevant to consider different groups of universities according to selectivity and other characteristics, in order to assess more precisely the effect of the AC items in their study programs. This may yield information regarding the usefulness of the items in the way they are built now and shed light regarding the construction of AC items for such study programs selection process.

The committee recommends that the mathematics test in PSU should be divided, and proposes the introduction of two new tests: A Basic math test and an Advanced math test. Since the PSU is constructed based on a curricular framework, each population of students must be assessed with its own curriculum; otherwise, important issues of learning opportunities and fairness are raised. In consequence, it is a sensible decision to have a mandatory test that considers only contents from the National Curriculum until the second high-school year, which corresponds to the common curriculum for all students. This issue is particularly important in mathematics where there are significant challenges in students' performance.

[3]For the analysis of predictive validity of DEMRE (2017), only the 45,530 students that graduated from secondary school in 2012 and who entered universities of the Council of Rectors and private universities participating in the SUA in 2013 were considered. Of these, 34,552 entered career programs with 30 or more students, which constituted the sample for the multiple regression analysis. No correction for range restriction was introduced. In SUA (2018), career programs with more than 25 students are considered, and the sample of students considered not only those who graduated in 2012 (SUA, 2018, p. 21).

Conceptualizing and constructing two new math tests, one Basic and one Advanced, may be highly beneficial for selection. BC and AC may be better assessed with two different tests that can be significantly shorter (as opposed to the current test comprised of 75 items), which may positively affect performance. Also, we envision that the Basic test may include contents from the curriculum for 7thand 8thgrades of primary school, providing better opportunities to start assessing the mathematical skills defined in the curriculum, which in theory must have a higher correlation with the performance in tertiary education programs, for which curricula are developed mostly according to competence-based models.

Another issue important to address is the modality in which these tests can be taken by students. For instance, one recent proposal of DEMRE is to have the Basic and the Advanced math test, and that each student could choose which one to take. (Both tests, however, could not be taken). This design may be appealing since the Advanced math test ends up being mandatory for some students and it is less time and resource consuming since students choose only one test. However, there are some important considerations to take into account under this design: to have only one test as mandatory would considerably increase its length and scope, and probably increase its difficulty significantly. Besides, if the high-performing students do not take the Basic math test, the test may end up with a lower score distribution, which could translate eventually into it becoming a low-difficulty test, sending a negative message to the school system. The choice between both tests may lead to some groups of students with greater risk aversion (especially females as shown in Arias**,** Mizala & Meneses, 2017) to choose the Basic math test, which might further decrease enrollment of specific groups of students in math-intensive study programs.

The committee identifies two different alternatives regarding the AC test. The first alternative is to make this test mandatory for all students applying to math-intensive study programs. This alternative has the disadvantage that many schools, particularly those with lower resources, might not teach advanced math curriculum, and therefore, students from those schools would have fewer learning opportunities. This alternative may also result in a more segregated applicant pool and student body in higher education programs. The second alternative is that both the BC and the AC are

mandatory. The more advanced test would not be relevant for a number of institutions and programs, but at least students would have the opportunity to show their knowledge and schools would not be discouraged from teaching advanced math content. The choice between the two alternatives is not obvious, and it requires careful consideration.

Proposal 4. To review the language test contents and particularly the assessment of writing in this test

The subjects of the Language and Communication section of the PSU are organized around two elements; processes and themes. Processes are generic aspects of the curriculum being evaluated, while themes refer to specific aspects of generic categories of texts. There are two writing processes that the test attempts to measure: i) to value writing as a creative and reflective activity of personal expression that allows one to organize ideas, present information, and interact with diverse realities and that provides an opportunity to elaborate a personal world vision consciously; and ii) the adequate use of a varied vocabulary, incorporating some less frequent uses and nuances of words, expressions and terminology according to the content, purpose and audience.

DEMRE's *Subjects of Language and Communication Test* states that these processes are measured indirectly. The first section is multiple choice dealing with handling connectors. In this section, students must select one of five connectors for ten sentences for which no context is provided. In the second part, called Writing Plan, the students must order a sequence of five statements to compose a coherent text. The task centers on measuring the capacity of the student to order a text already written. This way of evaluating writing is based on some assumptions that must be explained in detail and analyzed. First, it assumes that writing can be reduced to connectors and vocabulary. Second, it assumes that writing is general, i.e. that one way of writing serves to communicate different topics, and thus writing does not have to be situated or meaningful. Third, it assumes that it is not necessary to evaluate the production of writing directly and that it can be evaluated indirectly (Navarro, Ávila Reyes & Gómez Vera, 2018).

This method of evaluation only considers the most visible dimensions of writing—orthography, punctuation and grammar. However, these dimensions are not sufficient to evaluate if a text communicates ideas successfully, which is a fundamental function of writing (Perelman, 2018). Furthermore, this kind of evaluation requests tasks that are reproducible and can be copied and trained. By contrast, according to the experts, writing is situated, meaning that the activities involved are culturally and situationally

specific and changing. Writing is complex, since it includes multiple dimensions that require different knowledge and abilities; writing materializes complex processes of reflection (Navarro, 2018).

The situated and complex character of writing means that it cannot be evaluated as a generic, simple and transversal competence disconnected from the context, the audience and the specific purposes. Writing is materialized in texts or discursive formats that fulfill communicational functions in a social environment. In other words, adequate evaluation of writing requires asking for discursive formats that are pertinent and meaningful to students and that deal with specific topics with which they are familiar (Navarro, 2018).

The Importance of Writing

Writing is very relevant today since it is involved in most of the significant activities, tasks and events. This is especially true in formal education, given that it allows students to demonstrate what they have learned and to communicate their ideas. Writing is part of the integral formation of the students in higher education. It involves a number of the so-called 21st Century competencies, such as critical thinking, effective communication, argumentation, use and evaluation of information and sources, problem solving, creativity, etc. (Bazerman, 2013).

The wide use of writing in the knowledge society makes it necessary to incorporate writing in the instruments that measure the capacity to successfully complete a program of studies in an institution of higher education. Furthermore, developed countries include tests of writing in their education systems (Perelman, 2018).

Moreover, what is measured, and how it is measured, sends very strong signals to school systems. The result of using multiple choice in high-stakes standardized tests, such as the PSU, to evaluate writing is that the teaching and exercising of writing is not reinforced in high schools (Slomp, 2008). The deficit of this type of learning during high school has caused a number of Chilean higher education institutions to incorporate academic literacy initiatives (Preiss, Castillo, Grigorenko & Manzi, 2013a; Preiss, Castillo, Flotts & San Martín, 2013b). Including the evaluation of

writing in entrance tests for higher education could provide relevant input for these academic literacy initiatives.

This also relates to the validation of the tests in that validation involves the fact that the test is aligned with relevant educational objectives and that at least it does not limit the educational dominions that are taught (AERA, APA & NCME 2014).

How to score a written text

The instruments used to evaluate writing generally define an audience and a specific discursive genre explicitly (Perelman 2018). A written text may be evaluated with analytical or holistic rubrics. Analytic rubrics usually include 5 or 6 dimensions, such as content, orthography, vocabulary, cohesion, coherence and structure, each with 3-5 performance levels. One problem with this type of evaluation is how to quantify each dimension. It is also difficult to determine the weight of each dimension in the final evaluation and it is hard to achieve high reliability among analysts. In the case of the holistic rubric, the evaluation entails a more or less rapid reading; the focus is more on the strengths than on the weaknesses. Separate dimensions are not evaluated, but rather if the text is successful in communicating information and ideas to a pre-defined reader. It should ideally be evaluated by a panel, which increases the validity of the instrument in terms of fairness.

Chilean Experience in the Evaluation of Writing

The SIMCE test, which is aligned with the learning objectives of the curricular bases of the Ministry of Education, evaluates the writing ability of 6th graders specifying a purpose, a situation and a reader. What is evaluated preferably is the capacity to communicate a message with clarity. The test globally evaluates the written texts using a holistic guide for its evaluation.[4]

Apart from the SIMCE test, there are some pilot studies performed by DEMRE and a group of universities. DEMRE developed a research project to evaluate the possibility of implementing an instrument to directly assess writing in the university admission test. The researchers created the CODICE test which included multiple choice to evaluate reading comprehension and open questions based on the same text (with

[4]Complementarily, analytic guidelines are used to provide information to schools.

answers of medium length) to evaluate writing. Between 2006 and 2009, they implemented pilot tests and concluded that the inclusion of the open questions can enhance the university selection process. They also concluded that open questions referred to a given text can equilibrate differences in cultural conditions related to students' previous knowledge. Furthermore, they can lessen the impact of training for the test, as can happen with the use of an essay (Hernández, 2012).

The Pontificia Universidad Católica de Chile implemented a Written Communication Test to its undergraduate students. Students must write an essay choosing one out of three different topics; approximately 10 dimensions are evaluated using an analytic rubric. The test does not play a role in admission decisions. However, it is a high-stakes test since passing is a graduation requirement. If the students do not pass the test, they have to attend additional classes to improve their writing skills. Preiss et al. (2013a), using data from one cohort of students, found that this test is a significant predictor of university grades over time, even after controlling for individual and academic variables. Also, Preiss et al. (2013b), using similar data, found that the information originated by the writing test supplements information provided by the standardized PSU tests.

Moreover, a group of four universities participated in a pilot study in which instruments that complement the cognitive test used in the university admission process were developed and tested with high-school and university students. These instruments include a Critical Thinking Essay and a Personal Reflection Essay. The Critical Thinking Essay is based on the Written Communication Test mentioned above, the format of the test is the same, but the heading and also the instructions were modified. The study shows that these instruments can complement the current university admissions tests (mainly the language PSU test), although the Critical Thinking Essay has a relatively low reliability (Santelices et al., 2010).

International experience in the evaluation of writing

The National Assessment of Educational Progress (NAEP) evaluates the writing skills of a representative sample 4th, 8th and 12th grade students in the United States. The task provides a determined genre and recipient (for example, a letter to the school

Principal). During the test, different writing tasks must be completed in a given period of time. This test is evaluated using a holistic rubric.

The SAT program's optional writing test requires writing an essay in response to a passage. The essay must be completed in a specific amount of time.
The Third Comparative and Explicative Study (TERCE), developed by the Latin American Laboratory of Evaluation of the Quality of Education (LLECE), applies a test to a sample of 3rdand 6th graders from 15 countries of Latin America and the Caribbean. This test includes writing, using the letter format and a specific communicative purpose.

Conclusions and recommendations

There are good arguments for evaluating writing in tests that measure competences needed to successfully complete a program of studies in a higher education institution. Writing is part of a well-rounded education of students, since it allows them to communicate their ideas and knowledge and it is linked to the abilities needed to participate in the information, knowledge and communication society.

What is currently measured in the language part of the PSU battery is a proxy of writing based on a series of assumptions which are not fulfilled. The way writing is evaluated considers tasks that can be reproduced, copied and trained, which sends a message not to strengthen the teaching and exercising of writing in high schools.

There are good examples, even in Chile, of writing tests that fulfill a number of adequate requirements. However, their application implies greater monetary costs (since it is necessary to prepare people to score the test and ensure reliability) and correction time, which may be complex for the PSU given the short amount of time in which the tests results must be available.

Although in the short run it is not possible to implement an adequate evaluation of writing in the Language test of the PSU, the implementation of a writing test should be considered in order to signal schools the importance of teaching writing during the secondary education.

One possibility to evaluate is the introduction of a writing test that could be given during the 4th year of secondary education, since a test of this type would not require all the contents of secondary education, which allows a longer period of time for scoring. This test could be considered as proof of qualification for higher education and not part of the battery of tests that gives the entrance score to higher education. This is important since in general this type of tests have a lower degree of reliability. Some universities are already implementing this requirement from students who enter the first year and offer remedial programs to those who do not perform well. Adequate performance is a graduation requirement. The implementation of this writing test would send a message to educational institutions about the importance of teaching writing during high school.

For the time being, it is necessary to consider not including the sections of Connectors and Writing Plan in the Language test, since they do not measure students' writing competencies and they generate undesirable incentives for secondary education.

Proposal 5. To study the possibility of splitting the current Science test into separate Biology, Physics and Chemistry tests.

In this section, the issues of dimensionality and reliability of the PSU Science test are discussed. Both issues are relevant for the validity claims of the use of scores from the PSU Science test for admissions into higher education programs. Based on this discussion, recommendations regarding DEMRE´s proposal to split the current Science test into three separate tests are presented (DEMRE, 2017).

Internal Structure- Dimensionality

There are concerns about the reporting of one science score. The current version of the test includes (1) a basic section of 54 items from the three different Sciences (18 Biology items, 18 Chemistry items and 18 Physics items) and (2) an advanced section of 26 items from just one discipline (either Biology, Chemistry or Physics) for students from the scientific-humanistic track (DEMRE, 2016). Each student chooses which discipline to take based on his or her abilities and the admissions requirements of the program of interest. Students from the technical-professional track respond a different set of 26 items aligned with the professional track curriculum.

The question of whether the current test actually measures just one construct (unidimensional test) or multiple constructs (multidimensional test) directly relates to the way achievement on standardized tests should be reported. Single scores are used to report achievement on a unidimensional test, while scores on subscales are justified when tests measure more than one dimension. The dimensionality of a test is usually assessed using different sources of evidence that allow comparing the theoretical definition of the construct to the empirical evidence of the test´s internal structure (e.g., exploratory and confirmatory factor analysis).

In this case, the Science test combines questions from Biology, Chemistry and Physics, which may suggest that content-wise the test is not unidimensional. The evidence from factor analyses is somewhat contradictory (DEMRE, 2018a, 2018b). Although the results from confirmatory factor analyses performed by DEMRE staff show model fit indices that support a unidimensional structure, the results from an exploratory factor analyses show that the first factor of the Science test tends to

explain a proportion of the variance (between 9% and 25% depending on the test form) that is below the expected proportion of variance explained by just one factor in unidimensional tests [5]. This evidence is interpreted by committee members as suggestive of potential multidimensionality in the Science test and, therefore, the reporting of just one score is discouraged. Alternative ways of reporting achievement and/or ability should be considered.

Reliability

In addition to the dimensionality of the tests, it is important to take into consideration the precision of the assessment and the precision of the subscales in order to decide which scores and subscores should be reported. In other words, careful attention needs to be paid to the reliability of the test and its subdimensions in order to make a decision.

In the documents available to Committee members (DEMRE, 2017, page6), we see that the reliability of a "composite test score", comprised of all items from the basic and advanced science sections (a total of 44) from one specific discipline (either Biology, Chemistry or Physics), is within the range considered appropriate to report individual-level scores for the chosen discipline (0.90 to 0.92 depending on the science), as can be seen in the table below. No information on the reliability of the common basic Science items was found in the information made available to the Committee members.

Table. Reliability of Science Test and Composite Score

Confiabilidad (Alfa de Cronbach)	Prueba Ciencias (80 Preguntas)	Sección Materia Específica (44 preguntas)
PSU Ciencias -Biología	0.94	0.90
PSU Ciencias-Física	0.95	0.92
PSU Ciencias-Química	0.95	0.91

Source: DEMRE, (2017). *Propuesta del Departamento de Evaluación Medición y Registro Educacional (DEMRE) al Consejo de Rectores de las Universidades Chilenas (CRUCH) de*

[5]Stevens (1996) recommends a 75% or more for unidimensional tests, but Henson & Robertson (2006), in their review of exploratory factor analysis, report an average explained variance of 52% in unidimensional tests.

Mejoras a la PSU. Santiago: Chile.

Conclusions and recommendations

Committee members support DEMRE´s proposal to report the discipline of advanced content chosen by the student and the test score obtained by the examinee on the items of that specific discipline (combining basic and advanced items, i.e. "composite score"). This is actually one of the suggestions included in the Pearson Report of 2013.[6]

The Committee members disagree, however, on whether there was enough evidence to actually consider the Science test a "multidimensional" test. While some committee members were more inclined to consider the Science test as multidimensional, based mainly on exploratory factor analysis results, others were willing to consider the unidimensional structure as a possibility given the results from the confirmatory factor analyses. Even if considered a multidimensional test, there was not enough evidence to actually interpret and name the dimensions that the current version of the Science test may be measuring. Based on the discussion, committee members disagreed on what other additional Science score should be reported. While some supported the reporting of the score obtained in all 80 Science items, based on the results from the Confirmatory Factor Analyses, others thought the score on the basic items should be reported. The latter would only be acceptable provided that this section has an acceptable reliability, which is yet to be confirmed. According to the committee members endorsing this idea, the score on basic science items would be reported in addition to the "composite score" (the combination of basic and advanced items on the discipline chosen by each student).

It is important to consider that both the score on the basic science items and the "composite score" may be measuring a multidimensional construct. Further studies should be undertaken to explore this hypothesis.

[6]In this new scenario, it is important to ponder how the higher education admissions process will react to the availability of new information and how these decisions may impact school practices and students´ opportunities to learn the three Sciences. We believe that even if some programs decide to require specific "Advanced Sections" from applicants (for example, Medical schools may decide to ask for the Biology advanced section), all students would still need to learn all three Sciences, because there would still be a "Basic science test" that assesses all of them. Nevertheless, the effect of the potential requirement of certain "Advanced Sections" should be monitored closely to ensure that all high schools and students have the incentives to teach and learn Biology, Physics, and Chemistry.

The introduction of three separate Science tests requires further analysis and consideration. This modification may be based on the belief that it would improve the predictive validity of the PSU battery for some specific higher education programs, but that argument will require a close examination of the potential increases in predictive validity that may result as a consequence of the introduction of a “Basic” and an “Advanced” Math test. In addition, the number of potential users of these three separate Science tests should be closely estimated in order to examine whether there are enough potential users that justify the costs of developing, piloting and administering these three new national tests.

Proposal 6. To study the adoption of item response models to calculate scores and establish comparability in time across PSU tests.

Currently, the score on the PSU is a transformation of the number of items with correct responses. The document *Key Areas for Review in the PSU Test Battery*, dated April 2018, raises the issue of whether scoring methods based on item response theory (IRT) would be advantageous.

Making a transition to IRT–based scoring would have substantial implications in several areas. New software would be needed, along with new processes for item analysis, scoring, test assembly, equating, and quality control. Staffing requirements, including the needed training, would likely be affected as well. Making the transition would probably take 2 to 3 years and would require substantial advance research, including the investigation of competing IRT models and estimation procedures. Finally, publications would be needed in order to explain the change to the public, in particular, to high school students. This is a nontrivial issue, because, under the IRT models that would be useful for the PSU, scores would no longer have a simple relationship to the number of correct responses.

The item response models most widely used for items with dichotomous responses (scored incorrect or correct) are the one-parameter logistic model (1PL), the two-parameter logistic model (2PL) and the three-parameter logistic model (3PL). In these IRT models, the probability of correct response is represented as a function of the test-taker's proficiency level and the properties of the item.

The 1PL model considers only the item's difficulty, the 2PL also allows items to differ in discrimination (the degree to which the item can distinguish between high-proficiency and low-proficiency test takers) and the 3PL includes a third parameter representing the likelihood that a very low-proficiency test taker answers the item correctly (sometimes called the guessing parameter). Item responses are assumed to be unaffected by any other factors (such as item position or context) and are assumed to be independent, at a given proficiency level.

Using IRT-based scoring, however, can have substantial advantages. It can facilitate the equating of multiple test forms, which would make it more feasible to administer the PSU two or more times per year. It can also facilitate the creation of item banks and the process of assembling test forms. In particular, IRT makes it easier to construct tests that can accurately distinguish among test takers at particular points in the score scale. For example, if there is a particular cut score that is used in admissions decisions, IRT can simplify the development of a test that will yield precise scores in the vicinity of that cut score. A move to IRT would be essential if there were an interest in developing adaptive or multistage versions of some portions of the PSU. These types of tests adjust the difficulty of the test questions that are administered based on responses to previous questions (van der Linden & Glas, 2010; Yan, von Davier, & Lewis, 2014). It is important to note, however, that all these potential advantages of IRT apply only when the test data fit the IRT model and when there is enough data to obtain stable estimates of the model parameters. Therefore, tests of model fit are key to the decision of whether an IRT approach should be adopted, and if so, which model should be selected.

Currently, there is an interest in applying the Rasch model to PSU data. However, DEMRE's own analysis results, reported in images 5, 7, 9, 11, and 13 of the report *In reply to letter No. 19/2018 from SUA* show that few PSU items fit this model. This is not surprising since items on multiple-choice tests typically vary in discrimination and can be answered correctly by guessing. As an educational assessment expert noted 35 years ago, the assumptions of the Rasch model—that guessing is minimal and that item discriminations are equal—are not ordinarily met in the case of educational achievement data. These assumptions, in fact, "fly in the face of common sense and a wealth of empirical evidence accumulated over the last 80 years" (Traub, 1983, p.64). The Rasch model is appealing in its simplicity, but its attractive properties apply only if the data fit the model.

Below, we consider the scaling models used by prominent testing and assessment programs. Four of these are group-score assessments that are not high-stakes for individual test takers: the Programme for International Student Assessment (PISA), the National Assessment of Educational Progress (NAEP), the Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Reading

Literacy Study (PIRLS). Three are high-stakes admissions tests: the ACT, the SAT, and the Graduate Record Examinations (GRE). All seven of these programs either use the 2PL or 3PL models for multiple-choice items or use non-IRT approaches. It is significant that PISA previously used the Rasch model, but abandoned it because of its inadequacies.

PISA, NAEP, TIMSS, and PIRLS: According to PISA's 2015 technical report (Organisation for Economic Cooperation and Development, 2017, p. 142), "Concerns over the insufficiencies of the Rasch model to adequately address the complexity of the PISA data have been raised in the past" (Kreiner & Christensen, 2014; Oliveri & von Davier, 2011; among others). Other national and international studies utilize more general IRT models (Mazzeo & von Davier, 2014; von Davier & Sinharay, 2014). The National Assessment of Educational Progress (NAEP), for example, uses the three-parameter IRT model and the generalized partial credit model, or GPCM, (Allen, Donoghue & Schoeps, 2001) as does the Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Reading Literacy Study or PIRLS (Martin, Gregory & Stemler, 2000). To address the concerns about usage of the Rasch model, PISA 2015 implemented the two-parameter logistic model (2PLM; Birnbaum, 1968) for dichotomously scored responses and the generalized partial credit model (Muraki, 1992) for items with more than two ordered response categories.

ACT and SAT: These undergraduate admissions tests, used in the US, do not use IRT for score reporting. The reported scores are transformations of raw scores obtained through an equating process.

GRE: The GRE General Test is used for gaining admission to U.S. graduate programs, as well as some business schools and law schools. The current version of the GRE, which was launched in 2011, uses scales developed through item response theory. According to the GRE documentation by Robin & Steffen (2014):

> Two alternative IRT models were considered: the three-parameter model, which accounts for question difficulty, discrimination, and guessing (which may occur with questions that require test takers to select among a limited number of answer choices), and the two-parameter model, which accounts

> only for difficulty and discrimination. Analyses of both Verbal Reasoning and Quantitative Reasoning tryout data indicated that the IRT two-parameter model would provide a good fit to the data. This result was anticipated because the revised test greatly reduced the use of simple multiple-choice questions and, consequently, significantly reduced the possibility of successfully guessing answers. With fewer parameters to estimate, the two-parameter model requires smaller sample sizes than the three-parameter model previously used [by the GRE program]. Therefore, the two-parameter IRT model was chosen for the GRE revised General Test. (p. 3.3.4)

Conclusions and recommendations

Because of its far-ranging implications, the decision as to whether to adopt IRT scoring on the PSU requires careful consideration. In terms of model selection, DEMRE's own analyses suggest that adoption of the Rasch model would be highly risky. Using a model that does not provide a close fit to the data will yield inaccurate representations of item properties and test-taker skills. The Pearson Evaluation Report (2013), which recommends that the PSU make use of IRT models, suggests the use of the 3PL model (p. 62). The experiences of the GRE program, which elected to use the 2PL model in its most recent revision, are worth considering as well, should the PSU program decide to adopt IRT scoring. The evaluation of any proposed scoring procedure for the PSU should include the investigation of model fit, score precision, and the validity of scores as predictors of college performance.

Proposal 7. To appoint an International Technical Committee and a National Advisory Committee.

The committee agrees that it would be highly beneficial for the admission system to appoint an International Technical Committee composed by selected, world renowned experts that can provide advice on key matters and help build a strategic plan for the new admissions system. We commend DEMRE for appointing an International and National Advisory Committee and holding international conferences during these last years.

Summary of Recommendations

1) It is necessary to move forward in the development of the Theory of Action behind the University Selection Test, which would help make both the intended and unintended uses of the scores explicit.

2) It is recommended to separate the current Mathematics test into two separate tests: a basic test and an advanced test.

3) It is recommended to eliminate the indirect measurement of writing from the Language and Communication test and implement a direct measurement of writing in the short or medium term, which can be implemented after students complete 11th grade.

4) In the area of Science, it is recommended to report the discipline that students from the scientific-humanistic track choose and provide an additional score that reflects the performance in this area.

5) It is recommended to move forward in the development and implementation of a study plan that analyzes in depth the challenges that an eventual use of Item Response Theory would imply for the modeling, estimation, accuracy, reliability, validity and equating of achievement in the PSU, as well as the communication and understanding of PSU scores.

References

American Educational Research Association, American Psychological Association & National Council on Measurement in Education (AERA, APA & NCME) (2014). *Standards for educational and psychological testing*. Washington, D.C.: American Educational Research Association

Allen, N. L., Donoghue, J. R., & Schoeps, T. L. (2001). The NAEP 1998 technical report, NCES 2001-509, Office of Educational Research and Improvement, National Center for Education Statistics, U.S. Department of Education, Washington, DC: National Center for Education Statistics.

Arias, O. Mizala, A. and F. Meneses. (2017). "Brecha de género en matemáticas: El sesgo de las pruebas competitivas" mimeo CEA, Ingeniería Industrial, Universidad de Chile.

Bazerman, C. (2013) Understanding the Lifelong Journey of Writing Development. *Infancia y Aprendizaje: Journal for the Study of Education and Development 36*(4), 421-44.

Birnbaum, A. (1968). Some latent trait models and their use in inferring a student's ability. In F. M. Lord y M. R. Novick (eds.), *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

DEMRE. (2017). Evidencia preliminar para la reducción de contenidos en la PSU de matemáticas. Santiago: Chile.

DEMRE. (2018). Temario de Prueba de Lenguaje y Comunicación. Santiago: Chile.

DEMRE. (2018). Key areas for review in the PSU Test Battery. Santiago: Chile.

DEMRE. (2016). *Pruebas de Selección Universitaria. Informe Técnico, Volumén I, Características Principales y Composición.* Santiago: Chile.

DEMRE. (2017). Propuesta del Departamento de Evaluación Medición y Registro Educacional (DEMRE) al Consejo de Rectores de las Universidades Chilenas (CRUCH) de Mejoras a la PSU. Santiago: Chile.

DEMRE. (2018a). INFORME: Respuesta a carta N° 19/2018 del SUA. Santiago: Chile. Santiago: Chile.

DEMRE. (2018b). Follow up on the factor analysis results reported in Table 16 and 17. Santiago: Chile.

Donaldson, S. (2007). *Program theory-driven evaluation science: strategies and applications.* Nueva York, N.Y.: Lawrence Erlbaum Associates Publishers.

Donaldson, S. &Gooler, L. (2003). Theory-driven evaluation in action: lessons from a $20 million statewide work and health initiative. *Evaluation and Program Planning, 26*(4), 355–366. https://doi.org/10.1016/S0149-7189(03)00052-

Geiser, S., & Santelices, M.V. (2007). Validity of high school grades in predicting student success beyond the freshman year: School record vs. standardized tests as indicators of four-year college outcomes. Research & Occasional Paper Series CSHE.6.07. Center for Studies on Higher Education, University of California, Berkeley. Retrieved from http://cshe.berkeley.edu/ publications/docs/ROPS.GEISER._SAT_6.12.07.pdf.

Henson, R. K. & Robertsen, J. K. (2006) Use of Exploratory Factor Analysis in Published Research. Common Errors and Some Comment on Improved Practice. *Educational and Psychological Measurement, 66*(3).

Hernández, J. (2012) "Experiencia Prueba CODICE y el desarrollo de instrumentos complementarios para la selección universitaria" En Santelices, V., Ugarte JJ, y P. Kyllonen (eds), *Admisión a la Educación Superior: Mediciones Complementarias. Publicaciones Educación Superior Vol 1*, Ministerio de Educación.

Kane M. (2016). Explicating validity. *Assessment in Education: Principles, Policy & Practice, 23*(2), 198-211. https://doi.org/10.1080/0969594X.2015.1060192

Kreiner, S. & Christensen, K. B. (2014), Analyses of model fit and robustness: A new look at the PISA scaling model underlying ranking of countries according to reading literacy, *Psychometrika, 79* (2), 210-231.

Lacourly, N., San Martín, J., Amaya, J. & Uribe, P. (2017). INFORME DEMRE 2014-2016. Technical Report of the Center for Mathematical Modelling of Universidad de Chile.

Manzi, J., Bosch, A., Bravo, D., del Pino, G., Donoso, G. & Pizarro, R. (2010). Validez diferencial y sesgo en la predictividad de las pruebas de admisión a las universidades chilenas (PSU). *Revista Iberoamericana de Evaluación Educativa, 3*(2), 30-48. Retrieved from http://www.rinace.net/riee/numeros/vol3-num2/art2.pdf.

Martin, M. O., Gregory, K. D., &. Stemler, S. E. (Eds.). (2000). TIMSS 1999 Technical Report, International Study Center, Boston, MA.

Mattern, K., Patterson, B. & Kobrin, J. (2012). The Validity of SAT® Scores in Predicting First-Year Mathematics and English Grades. *Research Report 2012*-1. Retrieved from https://files.eric.ed.gov/fulltext/ED563105.pdf

Mazzeo, J. & von Davier, M. (2014), Linking scales in international large-scale assessments, In L. Rutkowski, M. von Davier and D. Rutkowski (eds.), *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis.* Boca Raton, FL: CRC Press.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm, *Applied Psychological Measurement, 16*(2), 159-177.

Navarro, F. (2018) "Más allá de la alfabetización académica: las funciones de la escritura en educación superior". In M.A. Alves & V. Iensen Bortoluzzi (eds.), *Formacao de profesores: Ensino, linguagens e tecnologías*. Porto Alegre,

Editora Fi.

Navarro, F., Ávila Reyes, N. & Gómez Vera, G. (2018) "Validez y justicia social: hacia la evaluación situada y significativa en pruebas estandarizadas de escritura". Working Paper.

NCME Statement, 2018. Recovered from https://higherlogicdownload.s3.amazonaws.com/NCME/c53581e4-9882-4137-987b-4475f6cb502a/UploadedImages/Documents/NCME_Position_Paper_on_Theories_of_Action_-_Final_July__2018.pdf

Oliveri, M. E. & von Davier, M. (2014). Toward increasing fairness in score scale calibrations employed in international large-scale assessments. *International Journal of Testing, 14*/1, pp. 1-21, doi:10.1080/15305058.2013.825265.

Organisation for Economic Cooperation and Development. (2017). PISA 2015 Technical Report. Retrieved from http://www.oecd.org/pisa/data/2015-technical-report/

Pearson. (2013). Final Report Evaluation of the Chile PSU.

Perelman, C. (2018) Towards a New NAPLAN: testing to the teaching activity. Surry Hills: NSW Teacher Federation.

Preiss, D., Castillo, J. C., Grigorenko, H. &Manzi, J. (2013a). "Argumentative writing and academic achievement: A longitudinal study". Learning and Individual Differences, 8, 204-211.

Preiss, D., Castillo, J. C., Flotts, P. & San Martin, E. (2013b). "Assessment of argumentative writing and critical thinking in higher education: Educational correlates and gender differences". *Learning and Individual Differences 28*, 193-203.

Robin, F., & Steffen, M. (2014). Test design for the GRE revised General Test. In Wendler, C. & Bridgeman, B. (Eds.), *The research foundation for the GRE revised General Test: A compendium of studies*, pp. 3.3.1-3.3.12. Princeton, NJ: Educational Testing Service.

Rogers, P., Petrosino, A., Huebner, R. & Hacsi, T. (2000). Program theory evaluation: practice, promise, and problems. *New Directions for Evaluation, 2000*(87), 5-13. https://doi.org/10.1002/ev.1177

Rossi, P., Freeman, H. & Lipsey, M. (2003). *Evaluation. A systemic approach.* Thousand Oaks, C.A.: Sage Publications.

Santelices, V. Ugarte, J., Flotts, P., Radovic, D., Catalán, X. & P. Kyllonen (2010) "Medición de atributos no cognitivos para el Sistema de Admisión a la Educación Superior en Chile". *Revista Iberoamericana de Evaluación Educativa 3*(2), 49-75.

Slomp, D. H. (2008) "Harming not helping: The impact of a Canadian standardized

writing assessment on curriculum and pedagogy". *Assessing Writing 12.*

Stevens, J. (1996) *Applied Multivariate Statistics for the Social Sciences* (3rd ed.) Mahwah, NJ: Lawrence Erlbaum.

Traub, R. E. (1983). A priori considerations in choosing an item response model. In R. K. Hambleton (Ed.), *Applications of item response theory*. Vancouver, BC: Educational Research Institute of British Columbia.

Van der Linden, W. J., &. Glas, C. A. W. (Eds.). *Elements of adaptive testing.* New York: Springer.

Von Davier, M. & Sinharay, S. (2014), Analytics in international large-scale assessments: Item response theory and population models. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis.* Boca Raton, FL: CRC Press,

Yan, D., von Davier, A. A., & Lewis, C. (Eds.). *Computerized multistage testing.* Boca Raton, FL: CRC Press.

Wikström, C., Wikström, M. (2017). Group differences in student performance in the selection to higher education: tests vs grades. *Frontiers in Education, 2*(45). https://doi.org/10.3389/feduc.2017.00045.

Zwick, R. (2017). *Who gets in? Strategies for fair and effective college admissions.* Cambridge, M.A.: Harvard University Press.

December 15, 2018

Appendix: Committee log of activities and communications

Membership: The active members of the Committee of Experts are Salomé Martinez, Alejandra Mizala, Veronica Santelices, and Rebecca Zwick. Below is a summary of the committee's meetings and communications with SUA.

Meetings:
June 22, 2018 Skype meeting: The participants were Salomé Martinez, Alejandra Mizala, Veronica Santelices, and Rebecca Zwick. The committee discussed its role and its response to the request dated June 13, 2018 from Maria Elena Gonzalez (see below).
September 24, 2018 Skype meeting: The participants were Alejandra Mizala, Veronica Santelices, and Rebecca Zwick. The committee discussed the report and determined who would draft each section.
November 26, 2018 Skype meeting: The participants were Salomé Martinez, Alejandra Mizala, Veronica Santelices, and Rebecca Zwick. The committee reviewed each section of the draft report.

Communications with SUA:
June 12: The English-language version of a SUA document, "Committee of Experts Unified Admissions System," dated May 31, 2018, was received from Tatiana Diener, proposing a timeline for the committee's work and a list of expected products.

June 13, 2018: A message was received from María Elena González, seeking committee responses to the May 31 document and indicating that requests for assistance could be addressed to Elisa Zenteno.

June 26, 2018: The committee sent a message to María Elena González saying, in part:
"We concluded that our first priority should be to address the seven issues described in the April 2018 document, 'Key Areas for Review in the PSU Test Battery.' We believe we can complete a report on these issues by the time of our next committee meeting in January 2019, assuming that we are able to obtain the materials we need."

June 29, 2018: The committee sent a communication to Elisa Zenteno requesting various reports and analyses.

July 5, 2018: The committee received a message from Elisa Zenteno asking for clarification of our message of June 26, 2018 and emphasizing the importance of issues concerning the students attending technical professional schools.

July 20, 2018: The committee sent a message to Elisa Zenteno and Maria Elena Gonzalez saying that we propose to consider the issues concerning students attending technical-professional schools in 2019.

July 23, 2018: The committee received a message from Elisa Zenteno agreeing that the committee could consider issues concerning students attending technical-professional schools in 2019 and letting us know she planned to send the documents we requested June 29.

July 30, 2018 Elisa Zenteno sent the committee the requested documentation in Spanish.

August 6-7. 2018: The committee exchanged communications with Elisa Zenteno regarding the contents of the documents sent on July 30 and the need to receive an English translation.

August 29, 2018: The committee received a message from Elisa Zenteno with the requested English translation.

October 1, 2018: The committee sent a message to Elisa Zenteno with technical questions about DEMRE documentation.

October 22, 2018: The committee's October 1 message was forwarded to Leonor Varas by María Elena Gonzalez.

November 9, 2018: The committee received a message from Maria Elena Gonzalez in which she forwarded a reply from DEMRE to the questions sent by the committee on October 1, 2018.

November 11, 2018: The committee received a message from María Elena Gonzalez with a report on students attending technical-professional schools.

November 13, 2018: The committee received a message from Maria Elena Gonzalez with a proposed agenda and schedule for further committee work.

November 26, 2018: The SUA´s proposed agenda and schedule were discussed. SUA proposed that only one member of the Committee travel to Arica to present the report on January 10th.The committee members agreed that all members should be present when discussing the Committee´s report and, therefore, it was not feasible to have that presentation on January 10th as suggested by SUA. Committee members agreed on submitting a report by December 20th for its editing and translation by SUA and to look for an alternative date in December to present the report to a group of chancellors in Santiago.

November 26 until December 11th: Veronica communicated with Tatiana Diener and Maria Elena in order to schedule the meeting. SUA staff members communicated that holding the meeting during December or the first fifteen days of January was not possible since most chancellors were busy with financial issues and end-of-year activities.

The main conclusions from the report were presented to a subset of academic authorities on January 28th.