



# Research Agenda

The IEA-ETS Research Institute (IERI) is a collaborative effort between the [Research & Development Division at ETS](#) and the [IEA](#) that will focus on improving the science of large-scale assessments. The IERI will undertake activities around three broad areas of work that include research studies related to the development and implementation of large-scale assessments

(research area); professional development and training (training area); and dissemination of research findings and information gathered through large-scale assessments (dissemination area).

## Goals of the Research Area

The IERI will undertake research activities that will address issues surrounding large scale assessments such as [IEA-TIMSS](#), [IEA-PIRLS](#), [OECD-PISA](#), [US-NAEP](#), IALS, ALL, etc.

The different research projects in this area collaborate by information technology as well as by face to face meetings, so that the area may be labeled a 'virtual' research center. Since the research projects will be hosted in the funding institutions, the exchange of ideas and findings will be facilitated by web-based collaboration as well as by the shared joint expertise of researchers involved in work on educational large scale assessments.

We expect that this virtual research area will contribute to the science of large-scale assessments so that the best available information is provided to policy makers and researchers from around the world.

The research area will focus on providing evidence in support of current statistical and psychometric methodologies, as well as developing new and improved methods. In addition, the research area will also focus on developing and validating constructs that could be used to predict and understand achievement results for policy relevant groups. More generally, the research activities will focus on providing research results that address issues that will help facilitate innovation and incremental improvement of large-scale assessment programs over time.

The research results from this area are expected to serve as the basis for operational innovations in future assessment cycles. Among other resources, the IEA and ETS will contribute the outcomes of their current research projects that are related to large-scale assessment. Results from the research activity are aimed at broadening the scope of topics and the base of researchers working on improving the methodologies involved in large-scale assessments in future assessment cycles.

## Research on Statistical Methodology

The IEA and ETS recognize that the statistical methods used in large scale educational surveys are an area of continuing importance. The statistical methods applied to report results within a country as well as cross-country comparisons have to be well grounded in statistical theory. At the same time, they need to be applicable and operationally feasible with relatively large and complex datasets. Policy makers and national coordinators have to be

well informed about the statistical methods applied and need to have access to documentation and tools for replicating results. The statistical methods applied in large-scale assessments are under increased scrutiny from experts all over the world and there is increased need to ensure that the best possible solutions given the existing constraints of the assessment cycle are chosen.

In order to ensure ongoing research on the statistical methodologies used in large-scale assessments, researchers affiliated with IEA and ETS will contribute a portion of their research time to further the development of statistical methods used in this international assessment.

The initial activities of the virtual research area will be concerned with projects that are related to the hosting institutions' own projects. National and international research funding agencies and foundations will be considered as funding sources for the research activities.

In addition to ongoing operational research on questions of immediate relevance for current assessment programs, longer term improvements to analysis tools and methodologies are of great importance to keep the assessment programs on the cutting edge of developments in psychometrics, sampling, and statistical analysis.

## **Areas Covered by Research Activities**

Projects undertaken by the virtual research area will tend to focus on one or more of the following 5 research priorities:

1. develop a more scientific approach to the development, use and interpretability of background questionnaires;
2. develop new constructs that extend the policy issues that might be addressed by these assessments;
3. improve the measurement of cognitive domains;
4. investigate effects of increased emphasis on the role of technology;
5. identify thematic issues to guide secondary analyses of existing data.

These 5 areas were selected given that most of the analysis can be carried out with existing data, the findings will inform future developments and provide direction to upcoming assessments, and the research will contribute to enhancing the quality and interpretability of the data. We hope that coordinating what is known and learned about large-scale assessments will ultimately enhance the visibility and utility of these assessments for policy makers.

## **Background Questionnaires**

Research on the areas and constructs reflected in the Background Questionnaires, Teacher Questionnaires, NRC and School Principal Questionnaires. Topics in this area include the development and application of methods to produce reliable derived variables, for example the use of latent trait and latent structure models on background questionnaire items that cover domains such as self-concept, information technology usage, or interests, time management, study skills etc. This area of research will focus on improvement of the relevancy of background questionnaires with respect to policy and decision making. It will also focus on developing an understanding of relation between the variables covered in the

background questionnaires and the cognitive outcomes. This area has both regional and international comparative aspects, so that some projects will be located in areas b) and e).

### **New Constructs**

Research will be supported to contribute to the selection of constructs and the improvement of the assessment of these constructs in the context questionnaires. The Center for New Constructs at ETS, which explores non-cognitive covariates of student's academic success has expressed an interest in participating in the virtual research lab. Among others, constructs such as time-management skills, self presentation and over-stating, as well as social learning are potential areas of interest and investigation. Researchers at ETS have developed a framework for non-cognitive assessment that can be used as an approach to studying the relation between non-cognitive variables and the cognitive domains assessed.

### **Cognitive Domains**

Research on cognitive constructs, assessment format, scoring, and modelling the cognitive domains are all areas of interest to the IERI. Potential topics in this area include the size and complexity of the conditioning models, scaling issues such as linking errors and item or block position effects. Alternative multidimensional models such as the M-IRT (Reckase, 1985; GDM, von Davier & Yamamoto, 2004; von Davier, 2005) etc. that can take into account testlet or passage effects, as well as position effects. Methodologies for the adjustment of results for exclusions, and mapping of state assessments to national or international scales may also be explored. Modelling of cognitive domains simultaneous with content domains is also of interest here. Studies addressing scoring and scoring reliability, the comparability of scoring across assessment cycles, and the application of measures that account for chance agreement in scoring reliability estimation are other areas of interest. (Hastedt??, Cohen, 1960; Batchelder & Klauer, 199x). Operational improvement and best practice studies that compare approaches used in different assessments with respect to their effectiveness to remediate scorer drift (e.g. NAEP) are quite important for maintaining and improving the quality of the data and also fall in this area of research topics.

### **Role of Technology**

Research on the potential and consequences of using information and communication technology to administer questionnaires hold great potential and are of continuing interest for the research area. Topics in this area include comparability of technology based and paper & pencil based questionnaires. Implications for future use on cognitive parts of the assessment, as well as feasibility and security issues.

### **Thematic Issues**

Research focussing on regions or specific countries, and aims at answering policy relevant questions will be facilitated. Research projects in this area may be requested and conducted using regional resources but may use results of research and data resources collected in other areas of research and other activities of the institute. Examples for regional studies are analyses that use national or international IEA data and address questions of concern for specific policy relevant groups (low income, immigrant populations, language minorities, ethnic minorities, gender effect, and interactions between such policy relevant groupings and cognitive as well as non-cognitive outcomes).

## **Additional Studies**

Research on the improvement of sampling processes, and modelling that takes into account the potential consequences of the complex sampling scheme. Topics include multilevel modelling, such as hierarchical linear models as well as methodologies for the correction of exclusions (Full Population estimates, FPE; McLachlan, 2006; Braun, 2006), and mapping of state assessments to national or international scales (Braun & Qian, 2005). Methods for estimating standard errors in complex samples are another topic of interest. Some of the studies in this area may simultaneously be located in area e) “regional studies”.

The current methodology used to analyze cognitive scales was developed by ACER to mirror developments of multilevel item response theory (IRT) models that originated at ETS to support the analyses of data for the National Assessment of Educational Progress (NAEP). ETS continues to fund an ongoing research agenda on extending and improving these methodologies. Von Davier, Sinharay, Oranje & Beaton (2006) delivered an overview of the current methodology and outlined steps for future extensions; see also Sinharay & Von Davier (2005) and Von Davier & Sinharay (2004) for detailed treatments of specific research questions around issues of high performance statistical computing for large datasets found in international assessments, and Von Davier (2003) for a comparison of the operational multilevel IRT used in practically all major educational survey assessments with a limited information approach.

Recent development of diagnostic models (Von Davier, DiBello & Yamamoto, 2006; Von Davier, 2005) have been applied to NAEP data (Xu & Von Davier, 2006) and promise to provide new ways to study the multidimensional structure of cognitive domains with this family of models. Another proposed area of research is the comparison of conditional and marginal approaches to estimating the item response parameters of the measurement model as input for the operational analyses.

The IERI will contribute research on mixed mode assessments. Like large-scale assessments, mixed mode assessments consist of questions with varying answer formats; e.g., multiple choice and open-ended questions. A statistical framework is being developed that provides solutions to problems surrounding these assessments such as how to handle data from multiple raters, missing data, and guessing. IEA will also explore the use of Generalizability Theory for large scale assessment programs and contribute funds for psychometric research around the ongoing large-scale assessments.

Multilevel extensions of the operational model (Fox & Glas, 2003) are another area of research of importance for the large-scale assessment program. The currently used approach can be described as a 2-level IRT model (Adams, Wilson & Wu, 1997). In order to incorporate school level effects appropriately, additional levels for the reporting analyses should be considered. Recent extensions on the basis of the General Diagnostic Model (GDM; von Davier, 2005) may be considered. Models of this family are non-parametric hierarchical model based on mixture distributions IRT models (see, for an overview, von Davier & Carstensen, 2007) which are based on hierarchical latent class models (Vermunt, 2003).

Projects may use psychometrics and multi-level modelling to improve these measures and interpretations. Effects of Schools or school types are usually expressed in terms of the proportion of total variance in educational achievement in a population of students, tied by the factor school; i.e. the between school variance. Projects targeting those types of are of

high interest for the lab, these could range from applications of hierarchical linear models using scaled achievement to the application of emerging models such as multidimensional multilevel IRT models. Current national and international data from a variety of large-scale assessment data will be used for these projects.

### **Use of Large-Scale Assessment Database**

The virtual research activity will be a focal point for professional development for national coordinators desiring training for secondary analysis as well as graduate students and emerging scholars wishing to do research using the large-scale assessment databases. ETS and the IEA will offer visiting scholar positions and summer internships, where research projects around the large-scale assessment databases could be located in the future. In addition, software tools to help external researchers can be provided together with trainings offered through the virtual research laboratory. IEA tools such as the IDB Analyzer software (GUI for jackknifing macros etc.), and ETS tools such as DESI (direct estimation software – interactive; Gladkova, Moran, Rogers & Blew, 2005) as well as research tools for recent model developments such as the software for estimating GDM's, mdltn (multidimensional discrete latent trait models; von Davier, 2005) could be provided in the form of end user licenses for research purposes, together with training on how to use these tools for secondary analyses of large-scale assessment data. These tools have a track record of use operationally in large-scale assessments, or have been used to explore new ways to analyse large scale assessment data by researchers of the proposing consortium. As an example, DESI grew out of the recognition that proprietary research software used for estimation and reporting in many large scale programs needs to be converted to a tool available to participating countries to ensure transparency and enable countries to carry out additional analyses if desired.

### **Initial projects**

1. Application of hierarchical multidimensional IRT models to proficiency data from large-scale assessment (GDM discrete MIRT applied to TIMSS or PISA data, MvD & X. Xu) for profile scores of subgroups.
2. Development and analysis of non-cognitive measures in terms of comparability and predictive power with respect to cognitive measures.
3. Full-population estimates: Adaptation of the improved method developed by H. Braun and J. Qian, for the inclusion of excluded parts of populations in the estimation of distributions.
4. Analysis of change, trend studies, item treatments or item deletion etc.
5. Alternatives to the full conditioning on all background variables (scaling BQ data, being selective in choosing reporting variables)
6. Equating of state tests (randomly equivalent populations, matching students or schools) to TIMSS or PIRLS (IGLU) results.
7. Collect locally dependent items into polytomous sum scores (Verhelst & Verstralen, 1997, and A. Arora's dissertation: applied to full sets of TIMSS scales)

## Management of the Research Area

The Research area of the IERI will be managed by Matthias von Davier, in collaboration with Juliane Barth, from the IEA. They will manage the research activities within the IERI virtual activity and head the effort to establish and coordinate the initial research agenda. Dr. Von Davier has been conducting a variety of research studies at ETS on extensions of the operational model used in educational survey assessments such as IALS, NAEP, TIMSS, PIRLS, and ALLS. Ms. Barth has extensive experience at the IEA-DPC working on TIMSS and PIRLS.

## References

Adams, R.J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics*, 22(1), 47-76.

Fox, J.P., & Glas, C.A.W. (2003). Bayesian modeling of measurement error in predictor variables. *Psychometrika*, 68, 169-191.

Gladkova, L., Moran, R., Rogers, A. & Blew, T. (2005) DESI – direct estimation software interactive. Manual for the ETS software. ETS: Princeton, New Jersey.

Sinharay, S. & von Davier, M. (2005) Extension of the NAEP BGROUP program to higher dimensions. ETS Research Report RR-05-27.

Rubin, D. B. & Thomas, N.(2002) Using Parameter Expansion to Improve the Performance of the EM Algorithm for Multidimensional IRT Population-Survey Models. In: Boomsma, A. van Duijn, M. & Snijders, T.: *Essays in Item Response Theory*, Springer: New York.

Verhelst, N. D. and Glas, C. A. W. (1995). The generalized one parameter model: OPLM. In *Rasch Models: Their Foundations, Recent Developments and Applications* (Edited by G.H. Fischer and I. W. Molenaar). Springer, New York.

Verhelst, N. D., Glas, C. A. W. and Verstralen, H. H. F. M. (1993). OPLM: One Parameter Logistic model. Computer program and manual. Arnhem: Cito.

Verhelst, N. D., Glas, C. A. W. and Verstralen, H. H. F. M. (1997). Modeling Sums of Binary Items by the Partial Credit Model. Research Report 97-07: Arnhem: Cito.

von Davier, M. (2003) Comparing Conditional and Marginal Direct Estimation of Subgroup Distributions. Research Report RR-03-02. ETS: Princeton, NJ.

von Davier, M. (2005) A General Diagnostic Model Applied to Language Testing Data. ETS Research Report RR-05-16.

von Davier, M (submitted for publication) Hierarchical General Diagnostic Models. Submitted to ETS RR series.

Von Davier, M. & Carstensen, C. H. (2007) *Multivariate and Mixture Distribution Rasch Models*. New York: Springer.

von Davier, M. & DiBello, L., & Yamamoto, K. (2006) Reporting test outcomes using models for cognitive diagnostics. ETS Research Report RR-06-28. To appear in: Klieme, E, Leutner, D. (eds.) *Assessment of competencies in educational contexts*.

von Davier, M. & Sinharay, S. (2004) A stochastic EM algorithm for latent regression Models. ETS Research Report RR-04-34.

- von Davier, M. Sinharay, S., Oranje, A. & Beaton, A. (2006) Marginal Estimation of Population Characteristics: Recent Developments and Future Directions. In C.R. Rao and S. Sinharay (Eds.), Handbook of Statistics (Vol. 26): Psychometrics. Amsterdam: Elsevier.
- von Davier, M. & von Davier, A.A. (2004) A Unified Approach to IRT Scale Linkage and Scale Transformations. Research Report RR-04-09. ETS: Princeton, NJ
- von Davier, M. & Yamamoto, K. (2004, October) A Class of Models for Cognitive Diagnosis. Invited Lecture at the ETS Spearman Invitational Conference, Princeton, NJ.
- Xu, X. & von Davier, M. (2006) Cognitive Diagnosis for NAEP proficiency data. ETS Research Report, RR-06-08.