

Signatura: EC 2020/109/W.P.4  
Tema: 5  
Fecha: 22 de mayo de 2020  
Distribución: Pública  
Original: Inglés

**S**



Invertir en la población rural

## **Análisis de sensibilidad de la evaluación del impacto en el marco de la FIDA10 y consecuencias para la FIDA11**

### **Nota para los miembros del Comité de Evaluación**

#### Funcionarios de contacto:

#### Preguntas técnicas:

##### **Sara Savastano**

Directora  
División de Investigación y Evaluación del Impacto  
Tel.: (+39) 06 5459 2155  
Correo electrónico: s.savastano@ifad.org

##### **Alessandra Garbero**

Economista Superior  
Tel.: (+39) 06 5459 2458  
Correo electrónico: a.garbero@ifad.org

#### Envío de documentación:

##### **Deirdre Mc Grenra**

Jefa  
Oficina de Gobernanza Institucional y Relaciones con los Estados Miembros  
Tel.: (+39) 06 5459 2374  
Correo electrónico: gb@ifad.org

Comité de Evaluación — 109.º período de sesiones  
Roma, 19 de junio de 2020

---

**Para información**

## I. Antecedentes

1. En su 127.º período de sesiones, la Junta Ejecutiva examinó el informe de evaluación del impacto en el marco de la Consulta sobre la Décima Reposición de los Recursos del FIDA (FIDA10) y las observaciones formuladas al respecto por la Oficina de Evaluación Independiente (IOE). La Junta agradeció a la Dirección por haber llevado a cabo la evaluación y acogió con agrado los resultados. También señaló la necesidad de que la Dirección estudiara las limitaciones de la metodología actual y trabajara para mejorar la metodología en el futuro. En concreto, la Junta solicitó a la Dirección que realizara un examen *inter pares* de la metodología y la fortaleciera aún más con el apoyo de un experto externo, y que considerara la posibilidad de compartir el muestreo y la metodología para su examen antes de llevar a cabo la evaluación del impacto en el marco de la Undécima Reposición de los Recursos del FIDA (FIDA11).
2. A fin de cumplir lo solicitado, la Dirección contrató a un experto<sup>1</sup> con dos propósitos: evaluar la metodología empleada en la preparación del informe de evaluación del impacto en el marco de la FIDA10 para determinar si había algún sesgo en los resultados derivado de la selección de los proyectos para la muestra; y, en función de lo determinado en esa primera tarea, confirmar el enfoque de selección para la evaluación del impacto que se realizaría en el marco de la FIDA11.
3. A raíz de la magnitud y la complejidad de los datos de la evaluación del impacto relativa a la FIDA10, se realizaron análisis de sensibilidad detallados respecto del enfoque aplicado en dicha evaluación. En el apéndice del presente documento figuran los resultados de los análisis y la validación de la muestra relativa a la FIDA11.
4. El FIDA seleccionó los proyectos de la FIDA10 que se someterían a la evaluación de impacto utilizando el protocolo aprobado por la Junta como parte del Marco relativo a la Eficacia de la Labor de Desarrollo del FIDA<sup>2</sup>. Como ha indicado la Junta recientemente, se realizó además un análisis de sensibilidad de los proyectos seleccionados a fin de poner a prueba la solidez de la muestra. Este análisis demostró que cualquier posible sesgo en la selección era insignificante, por lo que los resultados presentados en dicho informe eran válidos. El valor añadido por la metodología empleada en el informe compensa el sesgo encontrado y no debería socavar la labor excepcional del FIDA en cuanto a la presentación de informes. La misma metodología se aplicará para validar la muestra relativa a la FIDA11. Asimismo, se confirmó la validez del enfoque utilizado por el FIDA para evaluar los informes que se presentan.
5. La IOE y la Junta Ejecutiva estuvieron en lo correcto al plantear la posibilidad de que existiera un sesgo en la muestra de proyectos seleccionados, una preocupación justificada que surge ante cualquier muestra y, en especial, cuando no es posible realizar un muestreo aleatorio para hacer una selección. Una enseñanza clave de los análisis de sensibilidad es que es necesario tener en cuenta la posibilidad de que haya un sesgo en la etapa de selección de los proyectos. Habida cuenta de que en la selección de proyectos para la evaluación del impacto de la FIDA11 se ha seguido el protocolo aprobado por la Junta, dichos resultados también se validarán mediante un análisis de sensibilidad. Para la Duodécima Reposición de los Recursos del FIDA (FIDA12), se estudiará la posibilidad de aplicar otras metodologías.
6. Las conclusiones de los análisis demuestran que la metodología empleada en la FIDA10 es válida. Además, el proceso de selección de las muestras, en el cual se aplica el protocolo del Marco relativo a la Eficacia de la Labor de Desarrollo del FIDA, resulta adecuado para futuras actividades de evaluación del impacto. El enfoque no entraña ningún riesgo para la reputación del FIDA.

<sup>1</sup> Stefano Gagliarducci, Profesor de Economía del Departamento de Economía y Finanzas de la Universidad de Roma Tor Vergata e Investigador Afiliado del Instituto Einaudi de Economía y Finanzas. El Profesor Gagliarducci ha publicado artículos en importantes revistas de economía y ha realizado investigaciones anteriores sobre el sesgo de publicación.

<sup>2</sup> Marco relativo a la Eficacia de la Labor de Desarrollo del FIDA – Junta Ejecutiva de diciembre de 2016 (párrafo 58).

## II. Resumen de las conclusiones

7. Se realizaron varios análisis sistemáticos para detectar la posible existencia de un sesgo en los resultados del informe institucional del impacto, en particular en relación con los proyectos ejecutados en el marco de la FIDA10 que se habían seleccionado para la muestra y el grado en que esos proyectos eran representativos de la cartera de proyectos completados durante la FIDA10. El FIDA ha venido realizando evaluaciones del impacto con el fin de demostrar tanto la rendición de cuentas como el aprendizaje, por lo que la representatividad, el rigor y la transparencia son principios básicos de la metodología.
8. A continuación, se resumen las conclusiones de los análisis:

- i) **El posible sesgo en la selección de los proyectos para la evaluación del impacto en el marco de la FIDA 10 es insignificante.** Se llevó a cabo una evaluación sistemática de la muestra utilizada para realizar la evaluación del impacto relativa a la FIDA10 con el fin de determinar la posible existencia de un sesgo de selección, así como la naturaleza, la tendencia y la magnitud de ese sesgo. Para ello, se analizaron minuciosamente todas las posibles variables que podrían haber incidido en la selección. Entre esas variables se incluyeron las calificaciones relativas a los resultados de la ejecución, en concreto, las calificaciones disponibles al momento de la selección (julio de 2016), las cuales son las únicas variables que podrían haber influido en la selección. Sobre un total de 107 proyectos finalizados durante la FIDA10 y 24 calificaciones de ejecución, se examinaron las diferencias que existían entre las calificaciones medias correspondientes a la muestra, la cual equivalía al 15 % de los proyectos (19 proyectos), y las calificaciones medias del resto de la cartera (88 proyectos) a fin de evaluar si eran significativas. Se determinó que no había sesgo con respecto a la gran mayoría de las variables y que la diferencia en las calificaciones medias solo era significativa en el caso de dos variables: i) la tasa de desembolso y ii) los fondos de contrapartida. Por ello, se efectuó una validación exhaustiva para determinar la posible existencia de un sesgo de selección en cuanto a estas dos variables.

A tal fin, se realizaron metanálisis de subgrupos para determinar si la magnitud del impacto, medida en las evaluaciones del impacto a nivel de los proyectos, guardaba relación con la categoría de calificación relativa a los resultados de la ejecución (entre 1 y 6) a la que correspondía cada proyecto. Esto fue fundamental para verificar si existía alguna relación entre la magnitud del impacto y las escalas de calificación. En otras palabras, cabría esperar que los proyectos con calificaciones satisfactorias en cuanto a los resultados sobre los desembolsos o los fondos de contrapartida presenten una mayor magnitud del impacto en comparación con los proyectos calificados como moderadamente satisfactorios o insatisfactorios, lo cual podría dar a entender una relación positiva entre los resultados y el impacto. Sin embargo, las conclusiones de la Dirección demostraron que no había una relación clara entre la categoría de las calificaciones y las estimaciones del impacto, especialmente en el caso de las calificaciones relativas a los resultados sobre los desembolsos.

El objetivo estratégico relacionado con el acceso a los mercados ilustra este punto: los proyectos calificados como satisfactorios o con mejores resultados en cuanto a los desembolsos (en concreto, los catalogados con los códigos 1 a 3) tuvieron el menor impacto (57 %), frente a los proyectos calificados como moderadamente satisfactorios (80 %) y los proyectos calificados como insatisfactorios, los cuales, por su parte, presentaron el mayor impacto (89 %). Esto también se dio en gran medida en los demás objetivos estratégicos. En cuanto al objetivo estratégico relacionado con la capacidad

productiva, y la otra variable de calificación relativa a los fondos de contrapartida, los proyectos calificados como moderadamente satisfactorios (a saber, los catalogados con el código 4) tuvieron un impacto del 18 %, frente a los proyectos calificados como insatisfactorios (clasificados con los códigos 1 a 3). A partir de este análisis, la Dirección llegó a la conclusión de que la tendencia de las calificaciones, es decir, la obtención de mejores resultados, no se correspondía con un mayor impacto.

- ii) **Las correcciones del sesgo de selección determinaron que el sesgo era mínimo, lo que demostró que los resultados presentados en el informe de evaluación del impacto en el marco de la FIDA10 eran válidos.** Como segundo paso, y a fin de validar aún más estos resultados, se emplearon otros dos métodos para evaluar si era necesario ajustar los resultados de los metanálisis teniendo en cuenta la posible existencia de un sesgo de selección. El primer método, de corrección del sesgo de selección mediante el modelo de Heckman, se utilizó para computar la probabilidad que tenía cada proyecto en particular de ser seleccionado para una evaluación del impacto en el marco de la FIDA10 y, en caso de haber sido seleccionado, para ajustar las estimaciones del impacto institucional en función de este sesgo. Esta probabilidad tiene en cuenta los factores que determinan la selección, como los resultados relativos a los desembolsos, los fondos de contrapartida y otras variables observables clave<sup>3</sup>. El segundo método, de "recorte y relleno"<sup>4</sup>, está basado en las publicaciones sobre el metanálisis y, en un primer momento, se consideraba para verificar la presencia de sesgos generados por la inclusión deliberada de determinados proyectos o estudios en el metanálisis. Una limitación conocida del método de "recorte y relleno" es que puede corregir un sesgo de publicación inexistente, subestimando las magnitudes de los efectos (en el apéndice del presente documento se presentan los resultados de este método, cuyo nivel de fiabilidad es bajo)<sup>5</sup>. Con arreglo al primer método, se determinó que el sesgo de selección solo era ligeramente significativo (del 10 %) respecto del objetivo estratégico relativo al acceso a los mercados. En este caso, las estimaciones del impacto corregidas mediante el método de Heckman indicaron que, *ceteris paribus*, el impacto estaba sobrestimado en menos del 15 %<sup>6</sup>.
- iii) **Los proyectos seleccionados para la muestra que se utilizará en la evaluación del impacto en el marco de la FIDA11 no darán lugar a un sesgo.** Por último, se realizó una validación de la muestra relativa a la FIDA11 empleando el mismo enfoque. A la luz de los análisis de sensibilidad realizados con respecto a la FIDA10, se aplicó una metodología similar para validar la muestra de evaluaciones del impacto relacionadas con la FIDA11 y evaluar la existencia de un sesgo de selección en el momento en que se escogieron los proyectos (julio de 2018), utilizando las calificaciones relativas a los resultados de la ejecución y otras características de la cartera de proyectos en su conjunto. Como parte del programa para la evaluación del impacto en el marco de la FIDA11, se seleccionaron 24 de 112 proyectos a fin de realizar una evaluación rigurosa del impacto, lo que equivale al 21,4 % de la cartera de proyectos, el 20,9 % del total de la financiación y el 25,6 % de la financiación del FIDA. Tal como se hizo para la validación relativa a la FIDA10, se analizaron 24 calificaciones de ejecución, así como varias otras características objetivas de la cartera (como el número de beneficiarios y variables relacionadas con la financiación) a fin de determinar la posible

<sup>3</sup> Este enfoque se combinó con la metarregresión y el metanálisis.

<sup>4</sup> El método de "recorte y relleno" es un instrumento frecuentemente utilizado para detectar el sesgo de publicación y realizar ajustes en consecuencia.

<sup>5</sup> Terrin, N., Ch. Schmid, J. Lau e I. Olkin (2003): "Adjusting for publication bias in the presence of heterogeneity". *Statistics in Medicine*, 22: 2113-2126.

<sup>6</sup> 2 % en el caso de la movilidad económica, 15 % en el caso de acceso a los mercados, 10 % en el caso de la producción y 6 % en el caso de la resiliencia. Los resultados sobre nutrición se mantienen sin cambios.

presencia de sesgo al momento de la selección. No se detectaron diferencias significativas desde el punto de vista estadístico entre las características objetivas y las calificaciones de ejecución medias correspondientes a la muestra de proyectos elegidos (24) y las correspondientes a los demás proyectos del universo (88 proyectos que culminarían durante la FIDA11). Solo tenían importancia estadística los resultados de los sistemas de seguimiento y evaluación. Estas conclusiones demuestran que tampoco hay un sesgo de selección en el caso de la muestra relacionada con la FIDA11.

- iv) En conclusión, y a la luz de estas validaciones estadísticas y análisis de sensibilidad, la Dirección está en condiciones de confirmar que no hay sesgo de selección en las muestras de proyectos seleccionados para la evaluación del impacto institucional en el marco de la FIDA10 ni de la FIDA11.

# Appendix: Peer review of IFAD10 Impact Assessment Methodology

---

Stefano Gagliarducci, University of Rome Tor Vergata and EIEF

*In collaboration with* Alessandra Garbero and Sara Savastano, IFAD

## Contents

1. Introduction .....	2
2. Background .....	2
3. Definitions.....	4
3.1. Representativeness .....	4
3.2. Sampling/selection bias .....	5
4. Descriptive analysis and selection bias.....	6
5. Strategies for addressing the selection bias .....	15
5.1. Modelling selection bias for impact assessment using observables features .	15
5.2. Publication bias.....	15
6. Results from the Sensitivity Analyses.....	16
6.1. Subgroup meta-analyses .....	16
6.2. Sample selection bias correction á la Heckman .....	16
7. Conclusions on IFAD10 .....	19
8. Implications for IFAD11 .....	20
8.1. IFAD11 Sample Validation.....	20
8.2. Conclusions for IFAD11.....	34
Annex I	
Annex II	
Annex III	
Annex IV	

## 1. Introduction

During the discussion of the IFAD10 Impact Assessment Report at the 127th Executive Board meeting on September 11, 2019, the Board recommended to conduct Sensitivity analyses to assess the robustness of the corporate impact estimates and verify the results. This recommendation was made in light of the comments received by the Independent Office of Evaluation (IOE) and other Stakeholders, indicating the possible presence of a bias in the meta-analysis estimates and projections, raising concerns around the credibility of the findings. Such bias concerned the choice of the IFAD10 sample of projects selected for an Impact Assessment (IA) and notably, that such projects are, according to IOE, not representative of the portfolio of projects completing during IFAD10.

IOE's argument was based on a descriptive analysis of the performance ratings<sup>7</sup> at completion (or project completion reports ratings, in brief PCR). Their conclusion was that the projects selected for an impact assessment during IFAD10, seemed to include a large percentage of higher performing ones, therefore "potentially" yielding "biased" estimates of impact and possibly implying an overly optimistic vision of IFAD10 aggregate impact performance.

Systematic sensitivity analyses were therefore conducted to assess whether bias existed, and then investigate its magnitude. This is justified on transparency grounds, and on the fact that IFAD strongly believes in demonstrating accountability and learning, through rigorous methods.

Broadly speaking, sensitivity analysis is a process that allows the analyst to prove that the findings from a meta-analysis are not dependent on arbitrary or unclear decisions. In practice, they are aimed at repeating the meta-analysis, substituting alternative decisions or ranges of values for decisions that were arbitrary or unclear. For example, if the eligibility of some studies in the meta-analysis is dubious because they do not contain full details or are not representative, sensitivity analysis may involve undertaking the meta-analysis twice: first, including all studies and second, only including those that are definitely known to be eligible/representative. In this context and, through a weighting procedure, sensitivity analysis address the robustness of the results to the explicit inclusion of selection bias into the estimates, whereby this bias is assumed to be originated by the inclusion of a large number of projects with high performance ratings at completion, in the sample of the IFAD10 IAs.

The document presents the results of these analyses and is structured as follows. Section 3 recapitulates the background of IFAD approach to corporate reporting as stated in the Development Effectiveness Framework. Section 4 first introduce some definitions, section 5 presents a descriptive analysis, section 6 a literature review on the possible strategies to address bias, section 7 the results from the sensitivity analyses, section 8 concludes on IFAD10 and section 9 presents some implications for IFAD 11 and the corresponding validation of the IFAD11 sample of impact assessments.

## 2. Background

IFAD carries out project-level impact assessments (IAs) on a selection of projects (about 15 per cent) that are representative of the portfolio, to be able to measure corporate impact or aggregate development effectiveness. The latter requires a methodology that can attribute IFAD impact at the corporate level, e.g. provide an estimate of aggregate impact for the corporate indicators laid out in the IFAD Strategic Framework 2016-2025. The approach used is systematic, comprehensive, transparent, and builds upon the

---

<sup>7</sup> Since 2005, in line with the practice adopted in many other International Financial Institutions (IFIs) and United Nations organizations, IOE uses a six-point subjective rating system (where 6 is the highest score and 1 the lowest score) to evaluate projects. In addition to reporting on performance based on the six-point rating scale, in 2007 IOE introduced the broad categories of "satisfactory" (rating coded 4 to 6) and "unsatisfactory" (rating coded 1 to 3) for reporting on performance across the various evaluation criteria.

IFAD9 Impact Assessment Initiative methodology as well as the IFAD10 Development Effectiveness Framework.

IFAD's Development Effectiveness Framework (DEF)<sup>8</sup>, approved by the Board in September 2016 lays out the selection protocol to assess projects suitability to undergo an impact assessment, specifying to the following criteria:

- (i) potential to learn lessons;
- (ii) feasibility of conducting a scientifically rigorous impact assessment;
- (iii) buy-in from the government and IFAD;
- (iv) the capacity of a project to represent IFAD's portfolio and
- (v) the relevance of the impact assessment for subsequent project phases.

A key factor of impact assessment, in addition to accountability, is learning; and learning needs to inform the design of new projects in the same country or elsewhere. This provides a public good for policymakers. Therefore, a major recommendation approved by the Board - in the Development Effectiveness Framework -- stated that "impact assessments should have been selected and structured to facilitate and maximize learning while recognizing the need for corporate reporting, and that an impact assessment agenda should be a multi-stakeholder and participatory process to ensure relevance" (IFAD, 2016<sup>9</sup> pag.1).

Consequently, projects selected for IFAD10 IAs had to both display the potential for learning (innovative approaches or a clear evidence gap), while maintaining feasibility and have buy-in from the government.

In order to allow for adherence to the IFAD10 selection protocol, a working group was created to ensure that the selected projects were representative of the portfolio and revealed gaps for additional assessments, with a view to gaining an understanding of how projects fit into the portfolio. The expectation was that selected projects would have ultimately reflected the thematic and regional coverage of IFAD projects.

This led to a participatory process, finalized in September 2016, whereby projects selected for impact assessments were chosen in collaboration with IFAD's regional divisions to maximise this learning criteria. The divisions provided a list of projects suitable for inclusion based on the criteria specified according to the selection protocol. Subsequently, an appraisal was done to determine the impact assessments' feasibility in consultation with the regional divisions and relevant country directors.

Concerning corporate-level impact, IFAD's methodology to estimate aggregate development effectiveness involves a two-steps procedure whereby a meta-analysis of individual project-level impact assessment estimates is conducted in the first stage to compute aggregate corporate impacts, and a projection is conducted in the second stage to extrapolate impacts to the rest of the portfolio and estimate number of people benefiting across the portfolio<sup>10</sup>.

<sup>8</sup> The DEF was developed based on the lessons learned from the experience in demonstrating impact as part of the IFAD9 Impact Assessment Initiative. See [EB 2016/119/R.12](#)

<sup>9</sup> International Fund for Agricultural Development (IFAD), 2016. Development Effectiveness Framework ([EB 2016/119/R.12](#)).

<sup>10</sup> As far as the projection approach is concerned, this refers to a methodology that allows the estimated impact to be extrapolated to the whole IFAD portfolio, in order to obtain an assessment of the number of people that have benefited from IFAD investments. The corporate impact is interpreted as percentage change gain in each of the Strategic Objectives (SOs) and on IFAD's overarching goal. To translate this into the number of beneficiaries who benefited from IFAD's investments, distributional assumptions are needed to extrapolate the corporate estimates to the universe of beneficiaries in the portfolio.. The IFAD10 projection universe includes 107 projects, and is defined as the total number of projects completing during the replenishment period (2016-2018). As the projection require estimates of beneficiaries reached across the whole universe, the additional challenge has been to aggregate the number of beneficiaries for the overall portfolio. The information on the number of beneficiaries in the IFAD10 portfolio can be extracted from project documentation and IFAD internal reporting systems. Projected beneficiaries impacted are calculated based on the number of actual beneficiaries belonging to the universe of 107 projects. The latter amount to around 65.3 million beneficiaries. At the basis of the extrapolation, there are two main assumptions. One concerns the distribution of impacts, where the assumption is that corporate impacts are normally distributed with means and standard errors corresponding to the ones estimated empirically while obtaining aggregate impact estimates from the 17 impact studies covering 19 projects (equivalent to 18 per cent of the universe, actually). The second assumption is about defining what benefiting means in terms of exceeding a certain threshold. The projected number of beneficiaries impacted by IFAD's investments can be obtained by setting a threshold of at least 20 per cent for impact gains.



The aggregation is systematically done via a meta-analysis, a statistical procedure for combining data from multiple studies. Meta-analysis was pioneered in medical studies in the late seventies and then exponentially applied to clinical research. The meta-analysis is a study design used to systematically assess previous research studies to derive conclusions about a specific drug/treatment/research (or in our case policy) question. Outcomes from a meta-analysis may include a more precise estimate of the effect of treatment or risk factor for disease, or other outcomes (Haidich, 2010)<sup>11</sup>. More broadly, meta-analysis is defined as “the statistical analysis of a large collection of results for the purpose of integrating the findings” (Glass 1976<sup>12</sup>). In other words, it is “a quantitative summary of statistical indicators reported in similar empirical studies” (Brander et al. 2006<sup>13</sup>).

In the context of IFAD10 IA, the meta-analysis is a statistical procedure that aggregate the results of the 15% of projects on which an individual study is conducted. The outcome of the analysis is a proxy for an average effect (the treatment effect, or effect size) of the impact of IFAD’s overall portfolio. Once aggregated, corporate impacts were computed as percentage changes over the comparison group for each Strategic Objective (SO), notably production, market access or participation, and resilience, and for the overall IFAD goal of increased economic mobility.

### 3. Definitions

In this section, some definitions that are going to be useful for an understanding of the remainder of the document are provided.

#### 3.1 Representativeness

First, the concepts of representativeness, population or universe, and sample is defined: a representative sample is one that matches some characteristic of the underlying population, usually the characteristic that one is targeting with the research. In the context of IFAD10, the population refers to the population of projects that are in the universe of projects completing during IFAD10 (around 107 projects). The sample under analysis is defined as the 19 projects chosen for an impact assessment study.

As mentioned before, the IFAD10 selection protocol for the 19 IFAD10 Projects was based on a number of criteria to ensure representativeness of the portfolio. To what extent such criteria were as good as random is open to question. They could be quasi-random, in the sense that ex-ante, or at the time of selection, it was not possible to ascertain all of them.

In practice, it is almost impossible to ensure randomness due to a number of factors, such as feasibility of the impact assessment itself, knowledge asymmetries, political considerations and stakeholders buy-in, among others. It is worth recalling that the issue of representativeness of the impact assessment sample was also raised during the

---

Using estimates on the aggregate impacts and knowledge of the portfolio, one can then obtain projected number of beneficiaries benefiting above a 20 per cent threshold. In summary, projected beneficiaries impacted are obtained by randomly drawing a normal distribution of impacts with means and standard errors centred to the ones empirically estimated from aggregate impact distributions, thereby assuming that benefits are randomly and normally distributed and are above a specific threshold.

<sup>11</sup> “Important medical questions are typically studied more than once, often by different research teams in different locations. In many instances, the results of these multiple small studies of an issue are diverse and conflicting, which makes the clinical decision-making difficult. The need to arrive at decisions affecting clinical practice fostered the momentum toward “evidence-based medicine. Evidence-based medicine may be defined as the systematic, quantitative, preferentially experimental approach to obtaining and using medical information.” Haidich AB, 2010, *Meta-analysis in medical research*. Hippokratia. 2010 Dec.14 (Suppl 1):29-37.

<sup>12</sup> Glass, G. (1976). Primary, Secondary, and Meta-Analysis of Research. *Educational Researcher*, 5(10), 3-8.

<sup>13</sup> Brander L.M., et al. (2007). The recreational value of coral reefs: A meta-analysis, *Ecological Economics*, Vol. 63, Issue 1, 2007, 209-218.

previous replenishment cycle (IFAD9). Maintaining the integrity of the random selection conducted during IFAD9, was extremely difficult due to the above-mentioned factors. In that instance, projects were selected according to a number of criteria: 1) feasibility (suitable for an ex post impact assessment); 2) with the overarching aim of measuring the poverty reduction impact and, 3) statistically representative of the portfolio of activities undertaken by IFAD during IFAD9.

Therefore a representative sample of projects to be evaluated was determined by drawing a stratified random sample (a total of 41 projects, i.e. 26 first-choices and 15 reserves) from the universe of projects (with available datasets) closing between 2010 and 2015.

However, maintaining the integrity of the random sample proved difficult, as some randomly selected ones had to be replaced owing to both political and practical concerns (conflict setting, absence of PMU or key informants essential to gather retrospective information about projects, and impossibility to determine a counterfactual, among others). An internal consultation (in 2012) with IFAD Regional Directors and divisional representatives was then conducted to endorse the list of randomly selected projects to be evaluated by the external research partners. This process led to the replacement of 11 randomly selected projects with a set of purposively selected ones (purposive evaluations), given their strategic relevance and overall performance across the portfolio. Two of the purposively selected projects were dropped (namely, those in India and Senegal) after discussing the feasibility with internal staff. This last factor showed that even for “cherry picked” projects feasibility of an impact assessment was not guaranteed.

Notwithstanding these issues, IFAD sought to maintain the integrity of the representative sample and decided to maintain the randomly selected projects excluded from the final list of ex post evaluations and conduct the ex post assessments in-house with secondary datasets (14 Shallow dives).

Regarding IFAD10, and as noted above, a selection protocol was followed to ensure representativeness of the portfolio. The rationale for using a protocol is similar to what is normally conducted in the medical field, which is to randomly assign patients into treatment and control groups. As such, these protocols have features of quasi-randomness – as patients are selected into treatment – across a population of eligible patients<sup>14</sup>.

### 3.2 Sampling/selection bias

Selection bias is problematic because it is possible that a statistic computed of the sample is systematically erroneous. Selection bias can lead to a systematic over- or under-estimation of the corresponding parameter in the population. Selection bias occurs in practice as it is practically impossible to ensure perfect randomness in sampling (see before). If the degree of misrepresentation is small, then the sample can be treated as a reasonable approximation to a random sample. Also, if the sample does not differ markedly in the quantity being measured, then a biased sample can still be a reasonable estimate.

Selection bias is mostly classified as a subtype of selection bias, sometimes specifically termed sample selection bias, but some classify it as a separate type of bias. A distinction, albeit not universally accepted, of selection bias is that it undermines the external validity

---

<sup>14</sup> There is currently a heated debate around the topic of randomized controlled trials and whether they should be generally considered the gold standard, namely the best method to infer causality. It is worth recalling that in medical studies, researchers often choose not to randomize the intervention for one or more of the following reasons: (1) ethical considerations, (2) difficulty of randomizing subjects, (3) difficulty to randomize by locations (by region in the case of IFAD portfolio), (4) small available sample size (Harris et al. 2006).

of a test (the ability of its results to be generalized to the entire population), while selection bias mainly addresses internal validity for differences or similarities found in the sample at hand. In this sense, errors occurring in the process of gathering the sample or cohort cause selection bias, while errors in any process thereafter cause selection bias. In this specific case, it refers to selection bias.

However, selection bias and selection bias are often used synonymously.

## 4 Descriptive analysis and selection bias

In this section, a first assessment of the presence of selection bias is provided by presenting descriptive statistics that characterize the universe of IFAD10 projects, compared to the sample chosen for impact assessments. These descriptives are essential to understand the extent and the severity of the bias based on observable features. In presence of a selection bias, one should expect the two groups to differ significantly over an array of observable dimensions.

Recall that the sample is made of 19 projects, and that the universe is composed of 107 IFAD10 projects slated to close during IFAD10 at the time of selection. After projects were selected for assessment, a subset of the selected projects were extended such that their closing dates are now in IFAD11<sup>15</sup>.

The main argument against lack of representativeness cited by IOE was that the sample of IFAD10 IAs included a large majority of high performing projects as displayed by IOE's analyses of performance indicators at completion (PCR ratings).

Thus, Table 1 reproduces the one presented by IOE in their summary document for the Evaluation Committee Session (EC) held in September 2019. Average performance indicators at completion (PCR ratings) are displayed for the sample of 19 projects evaluated as part of IFAD10 IAs and the remaining 88 projects not evaluated out of the total universe of 107 projects.

PCR ratings at completion are subjective ratings with a six-point measurement scale system ranging from 6 to 1, with 1 being the lowest score across each criterion. At the time of IFAD10 selection, when the sample of 15% of projects was identified, none of these scores were available for consultation nor officially available within IFAD official system<sup>16</sup>.

While comparing the two tables, a number of issues became apparent. The first, is the lack of definition of the universe of projects analysed e.g. the total number of observations (projects), in IOE's document. As a consequence statistics for unselected projects completing during IFAD10 (columns 3 and 4) varied across some indicators, while results for the IFAD10 sample (namely columns 1 and 2) coincided with IOE calculations. Also, PCR scores were not available for all the unselected projects, therefore the total number of observations by indicator varies between 64 to 88 projects (column 3). Last, and similar

---

<sup>15</sup> Notably Bangladesh (CCRIP), Kenya ( SCDP), Sao Tome (PAPAC), Rwanda (PRICE) will now close in IFAD11.

<sup>16</sup> Specifically, regarding features that might have driven the IFAD10 IAs selection process, and alter the representativeness of the sample of the IFAD10 projects portfolio, the following ones were available at the time of the selection, notably in 2016: the project type or sector, the region of implementation, the size of outreach, the disbursement performance, and the implementation performance indicators. As projects were ongoing, project completion report ratings (the one verified by IOE) were not available to inform the selection.

to IOE's table, final PCR ratings are only available for 13 of the 19 projects that underwent an IA. The unavailability of the PCR ratings for the whole IA sample, is due to the fact that PCRs cannot be finalized if projects completion dates are extended. This was the case of 6 projects out of 19, whereby their completion dates were extended into IFAD11.

Therefore, Table 1 shows the difference across PCR ratings as presented by IOE in their comments. T-tests were run for the statistical significance of the difference in means (balance tests).

Before commenting the table, it is important to highlight that, while this is certainly an informative exercise, the latter should be taken with caution. As stated in the DEF, ex-post impact assessment should ideally occur prior to the closure of the project, so project completion reports can benefit from the impact assessment findings. If so, PCRs ratings incorporate IA findings when available – hence potentially influencing the direction of the final rating. Therefore, from a statistical standpoint, PCR ratings should not be used to assess the presence of selection bias, as they are positively affected by the mere virtue of a project being under evaluation.

**Table 1: Balance tests: PCR ratings**

	Average PCR ratings (IFAD10 IA sample)		Average PCR ratings ( <b>completing</b> IFAD10 projects 2016-2018)		Sample - Unselected	
	(1) N. projects	(2) Mean	(3) N. projects	(4) Mean	(5) Diff. in Means	(6) P-score
<b>Relevance</b>	13	5.2	75	4.6	0.6	0.005
<b>Effectiveness</b>	13	4.8	76	4.2	0.6	0.005
<b>Efficiency</b>	13	4.4	76	3.8	0.6	0.026
<b>Sustainability</b>	13	4.4	76	3.8	0.5	0.017
<b>Project performance</b>	13	4.7	75	4.1	0.6	<b>0.002</b>
<b>Rural poverty impact</b>	13	4.8	75	4.1	0.7	0.001
<b>Gender equality and women's empowerment</b>	13	4.6	88	0.6	0.2	0.38
<b>Innovation</b>	13	4.8	76	4.4	0.4	0.111
<b>Scaling up</b>	13	4.8	75	4.4	0.5	0.066
<b>Environment and natural resource management</b>	13	4.5	73	4.1	0.4	0.056
<b>Adaptation to Climate Change</b>	11	4.5	64	4.1	0.5	0.022
<b>IFAD performance</b>	13	4.8	76	4.3	0.5	0.01
<b>Government performance</b>	13	4.7	76	4.1	0.6	0.014
<b>Overall project achievement</b>	13	4.8	75	4.2	0.6	0.004

Source: Calculations based on IFAD10 IA sample and data extracted from IOE ratings database.

Table 1 shows average subjective rating scores across 14 mandatory criteria<sup>17</sup>, used by IOE to evaluate projects at completion. However, means of selected and unselected projects are based on the universe of projects as defined by Management in the IFAD10 Report (107 completed projects).

Given the above concern with the use of PCR ratings, in what follows, the significance of differences in pre-determined characteristics is tested. These essentially are baseline characteristics and include objective features and 24 implementation ratings as measured at the beginning of the project (i.e., the first indicator of performance that is available in the system). Implementation ratings are monitored during the lifespan of the project. These are the ones that, effectively, should have informed projects' selection at the beginning of the IA process. Note that while the first three indicators in the table are objective, notably project duration, number of beneficiaries and total approved funding, performance indicators are self-assessed and are expressed on a rating scale (1-6) ranging from unsatisfactory, to highly satisfactory<sup>18</sup>.

Zooming in, note how the projects selected for impact assessments were similar on average in terms of financing and number of actual beneficiaries to the universe of projects. The average approved financing across the sample of IAs was \$51.7 million, and the average in the universe was of \$50.9 million. In terms of beneficiaries, the average number of beneficiaries was 610,556 in the universe and 490,339 in the IA sample, but this difference is not statistically significant. In almost all performance ratings categories, the IAs performed slightly better than the universe of projects, on average. However it is

---

<sup>17</sup> Based on IOE Manual (2015) pp. 38 -40. These definitions build on the OECD/DAC Glossary of Key Terms in Evaluation and Results-Based Management; the Methodological Framework for Project Evaluation agreed with the Evaluation Committee in September 2003; the first edition of the Evaluation Manual discussed with the Evaluation Committee in December 2008; and further discussions with the Evaluation Committee in November 2010 on IOE's evaluation criteria and key questions. Rural poverty impact is defined as the changes that have occurred or are expected to occur in the lives of the rural poor (whether positive or negative, direct or indirect, intended or unintended) as a result of development interventions. Project performance is an average of the ratings for relevance, effectiveness, efficiency and sustainability of benefits. Relevance measures the extent to which the objectives of a development intervention are consistent with beneficiaries' requirements, country needs, institutional priorities and partner and donor policies. It also entails an assessment of project design and coherence in achieving its objectives. An assessment should also be made of whether objectives and design address inequality, for example, by assessing the relevance of targeting strategies adopted. Effectiveness is the extent to which the development intervention's objectives were achieved, or are expected to be achieved, taking into account their relative importance. Efficiency is a measure of how economically resources/inputs (funds, expertise, time, etc.) are converted into results. Sustainability of benefits (or simply sustainability) is the likely continuation of net benefits from a development intervention beyond the phase of external funding support. It also includes an assessment of the likelihood that actual and anticipated results will be resilient to risks beyond the project's life. Gender equality and women's empowerment measures the extent to which IFAD interventions have contributed to better gender equality and women's empowerment, for example, in terms of women's access to and ownership of assets, resources and services; participation in decision making; work load balance and impact on women's incomes Nutrition and livelihoods. Innovation and scaling up (OR scaling up) measures the extent to which IFAD development interventions: (i) have introduced innovative approaches to rural poverty reduction; and (ii) have been (or are likely to be) scaled up by government authorities, donor organizations, the private sector and others agencies. Environment and natural resource management represents the extent to which IFAD development interventions contribute to resilient livelihoods and ecosystems. The focus is on the use and management of the natural environment, including natural resources defined as raw materials used for socio-economic and cultural purposes, and ecosystems and biodiversity – with the goods and services they provide. Adaptation to climate change is the contribution of the project to reducing the negative impacts of climate change through dedicated adaptation or risk reduction measures. Performance of Partners (IFAD and Government): This criterion assesses the contribution of partners to project design, execution, monitoring and reporting, supervision and implementation support, and evaluation. The performance of each partner will be assessed on an individual basis with a view to the partner's expected role and responsibility in the project life cycle. Finally, overall project achievement provides an overarching assessment of the intervention, drawing upon the analysis and ratings for rural poverty impact, relevance, effectiveness, efficiency, sustainability of benefits, gender equality and women's empowerment, innovation and scaling up, as well as environment and natural resources management, and adaptation to climate change.

<sup>18</sup> Ratings of project performance should be consistent with the findings of progress reports and of the supervision mission report. By rating each indicator, different criteria are applied as explained below, however in general the ratings are:

(6) Highly satisfactory. Targets/requirements met or exceeded. Considered as best practice.

(5) Satisfactory. Targets/requirements met with only minor delays or set-backs.

(4) Moderately satisfactory. Most targets/ requirements met but delays or set-backs experienced.

(3) Moderately unsatisfactory. Some targets/ requirements met but issues/constraints have negatively affected implementation.

(2) Unsatisfactory. Few targets/requirements met. Issues/constraints remain unresolved. Delays have seriously undermined implementation.

(1) Highly unsatisfactory. Almost no targets/ requirements met. Consideration should be given to cancellation/suspension.

important to note that these differences are not statistically significant for the majority of indicators presented – except for the following ratings:

- Assessment of the Overall Implementation Performance\* : significant at 10% level.
- Acceptable Disbursement Rate\*\* : significant at 5% level.
- Counterparts Funds\*\* significant at 5% level.
- Coherence between AWPB and Implementation\* : significant at 10% level.

**Table 2: Balance tests: implementation performance ratings (baseline characteristics)**

	<i>IFAD10 IA Sample</i>		<i>IFAD10 projects</i>		<i>Unselected</i>	<i>Sample - Unselected</i>
	<b>N. projects (1)</b>	<b>Mean (2)</b>	<b>N. projects (3)</b>	<b>Mean (4)</b>	<b>Diff. in Means (5)</b>	<b>P-score (6)</b>
<b>Project Duration</b>	19	8.16	88	8.27	-0.12	0.863
<b>Beneficiaries</b>	19	490 339	88	636 512	-146 173	0.732
<b>Approved Funding</b>	19	51 712 292	88	50 700 000	986 098	0.957
<b>Assessment of the Overall Implementation Performance Likelihood of Achieving the Development Objective Effectiveness</b>	19	4.11	88	3.85	0.25	0.059
	19	4	88	3.98	0.02	0.857
	12	3.83	69	3.88	-0.05	0.74
<b>Targeting and Outreach</b>	19	4.26	88	4.11	0.15	0.224
<b>Gender equality &amp; women's participation</b>	19	4.05	88	3.99	0.06	0.666
<b>Agricultural Productivity</b>	15	4.13	71	3.94	0.19	0.147
<b>Adaptation to Climate Change</b>	2	4	11	3.91	0.09	0.863
<b>Institutions and Policy Engagement</b>	18	4.11	78	4.01	0.1	0.549
<b>Human and Social Capital and Empowerment</b>	15	4.13	77	3.92	0.21	0.177
<b>Quality of Beneficiary Participation</b>	19	3.95	88	4.06	-0.11	0.401
<b>Responsiveness of Service Providers</b>	19	3.89	88	3.97	-0.07	0.592
<b>Environment and Natural Resource Management</b>	2	4	13	3.77	0.23	0.607
<b>Exit Strategy</b>	11	4.09	58	3.97	0.13	0.387

<b>Potential for Scaling-up</b>	14	4.21	72	4.07	0.15	0.366
<b>Quality of Project Management Knowledge Management</b>	19	3.95	88	3.85	0.1	0.631
	16	4.19	76	4.03	0.16	0.243
<b>Coherence between AWPB and Implementation</b>	17	4.12	81	3.79	0.33	0.064
<b>Performance of M&amp;E System</b>	19	3.74	87	3.83	-0.09	0.574
<b>Acceptable Disbursement Rate</b>	19	4.21	88	3.43	0.78	0.028
<b>Quality of Financial Management</b>	17	4.12	79	3.9	0.22	0.218
<b>Quality and Timeliness of Audit</b>	19	4.11	87	4.01	0.09	0.582
<b>Counterparts Funds</b>	19	4.42	88	4.01	0.41	0.031
<b>Compliance with Loan Covenants</b>	19	4.21	88	4.02	0.19	0.189
<b>Procurement</b>	19	4.16	88	4	0.16	0.292

Table 3 and Table 4 show the distribution of the sample of IAs projects by IFAD's region and project sector or type. In the universe, 30 projects were in the Asia and Pacific Region (APR) followed by 26 in Western and Central Africa (WCA), 20 in Eastern and Southern Africa (ESA), 18 in Latin America and Caribbean (LAC), and 13 in the North East and Northern Africa region (NEN). As far as the IAs Projects' Sample is concerned, the majority of IAs (six) were conducted in ESA, while five were conducted in APR, four in WCA, three in LAC, and one in NEN. Table 22 in the Annex presents the mean performance by region and shows similar results i.e. that none of the mean ratings are statistically different across the IA sample and the unselected projects, although there is more variation, largely due to the lower number of overall projects with each region.

Turning to the project sector or type, a variable that is quite broad in the current classification system, the majority of projects in the universe are classified as agricultural development (37), rural development (34), and credit (14). However, no credit projects were selected for assessment in IFAD10 and over 40% of all IAs were of rural development projects. Nevertheless, because the project sector categorization is extremely broad, contains considerable overlap among categories, and is insufficiently informative about the true nature of the project, it lacks utility for the conduction of rigorous sensitivity analysis or bias estimation.

**Table 3: Distribution of Projects in the Universe and in the IAs sample by Region**

<b>Universe by Region</b>			<b>IAs by Region</b>		
<b>BU</b>	<b>Projects</b>	<b>%</b>	<b>BU</b>	<b>Projects</b>	<b>%</b>
APR	30	28.04	APR	5	26.32
ESA	20	18.69	ESA	6	31.58
LAC	18	16.82	LAC	3	15.79
NEN	13	12.15	NEN	1	5.26
WCA	26	24.30	WCA	4	21.05
Total	107	100.00	Total	19	100.00

**Table 4: Distribution of Projects in the Universe and in the IAs sample by Sector or Project type**

<b>Universe by Project</b>			<b>IAs by Project</b>		
<b>Sector</b>	<b>Projects</b>	<b>%</b>	<b>Sector</b>	<b>Projects</b>	<b>%</b>
AGRIC	37	34.58	AGRIC	6	31.58
CREDI	14	13.08	CREDI	0	0.00
FISH	2	1.87	FISH	0	0.00
IRRIG	7	6.54	IRRIG	1	5.26
LIVST	4	3.74	LIVST	2	10.53
MRKTG	6	5.61	MRKTG	1	5.26
RSRCH	3	2.80	RSRCH	1	5.26
RURAL	34	31.78	RURAL	8	42.11
Total	107	100.00	Total	19	100.00

Finally, implementation performance ratings were combined (Table 5) to assess and test for differences across proportions/percentage of projects rated satisfactory both in the IA sample(19) and in the universe of projects completing during IFAD10 (107). Note that there was not much variation in project scores in either the universe or the IA sample with



the highest density of projects around scores of four and five out of six. Specifically, when the indicators are transformed to indicate whether a project scored a satisfactory (4-6) or unsatisfactory (1-3) rating, it is apparent that in both the IFAD10 IA sample and the universe the majority of projects received satisfactory ratings. In the IA sample, between 84 and 100 percent of projects received satisfactory ratings and in the universe between 75 and 98 percent of projects did.

Although there are differences in the relative frequency of unsuccessful projects, it is clear that the majority of portfolio projects receive satisfactory scores. As such, it is reasonable that a high proportion of the sample would be high performing projects.

In summary, a conclusion that can be drawn from this analysis, is that bias is absent for the majority of baseline indicators presented – except for the following two ratings:

- Acceptable Disbursement Rate\*\*: significant at 5% level.
- Counterparts Funds\*\* significant at 5% level.

Further analyses are therefore conducted on these variables in question in the following sections.

**Table 5: Balance tests: Proportions of Projects Rated Satisfactory**

	<i>Average performance ratings</i>	<i>Average performance ratings</i>	<i>Difference in IAs and Universe Means</i>	<i>Proportion of IAs Rated</i>	<i>Proportion of Universe Rated</i>	<i>Sample- Universe</i>	<i>Significance</i>
	<i>IFAD10 IA Sample</i>	<i>(Completing IFAD10 projects 2016-2018 : universe(107))</i>		<i>Satisfactory</i>	<i>Satisfactory</i>		
	<b>Mean (1)</b>	<b>Mean (2)</b>	<b>Diff. in Means (3)</b>	<b>Proportion (4)</b>	<b>Proportion (5)</b>	<b>Diff. in Proportion (5)</b>	<b>P-score (6)</b>
<b>Assessment of the Overall Implementation Performance</b>	4.11	3.9	0.21	100%	92%	8%	0.436
<b>Likelihood of Achieving the Development Objective</b>	4	3.98	0.02	100%	84%	16%	0.561
<b>Effectiveness</b>	3.83	3.88	-0.04	100%	83%	17%	0.957
<b>Targeting and Outreach</b>	4.26	4.14	0.12	100%	98%	2%	0.336
<b>Gender equality &amp; women's participation</b>	4.05	4	0.05	89%	89%	1%	0.93
<b>Agricultural Productivity</b>	4.13	3.98	0.16	100%	84%	16%	0.569
<b>Adaptation to Climate Change</b>	4	3.92	0.08	94%	90%	5%	0.685
<b>Institutions and Policy Engagement</b>	4.11	4.03	0.08	100%	82%	18%	0.537
<b>Human and Social Capital and Empowerment</b>	4.13	3.96	0.18	95%	87%	8%	0.483
<b>Quality of Beneficiary Participation</b>	3.95	4.04	-0.09	100%	95%	5%	0.833

<b>Responsiveness of Service Providers</b>	3.89	3.95	-0.06	95%	87%	8%	0.221
<b>Environment and Natural Resource Management</b>	4	3.8	0.2	100%	87%	13%	0.582
<b>Exit Strategy</b>	4.09	3.99	0.11	84%	75%	9%	0.356
<b>Potential for Scaling-up</b>	4.21	4.09	0.12	100%	90%	10%	0.355
<b>Quality of Project Management</b>	3.95	3.87	0.08	84%	82%	2%	0.851
<b>Knowledge Management</b>	4.19	4.05	0.13	100%	85%	15%	0.293
<b>Coherence between AWPB and Implementation</b>	4.12	3.85	0.27	84%	69%	15%	0.317
<b>Performance of M&amp;E System</b>	3.74	3.81	-0.07	78%	72%	7%	0.993
<b>Acceptable Disbursement Rate</b>	4.21	3.57	0.64	95%	65%	29%	0.025
<b>Quality of Financial Management</b>	4.12	3.94	0.18	89%	79%	11%	0.764
<b>Quality and Timeliness of Audit</b>	4.11	4.03	0.08	89%	93%	-3%	0.486
<b>Counterparts Funds</b>	4.42	4.08	0.34	79%	74%	5%	0.194
<b>Compliance with Loan Covenants</b>	4.21	4.06	0.15	100%	87%	13%	0.492
<b>Procurement</b>	4.16	4.03	0.13	89%	74%	16%	0.681

## 5 Strategies for addressing the selection bias

Given that there are statistically significant differences in only two observable features of the projects subject to an impact assessment, two main strategies are considered to assess the need for adjusting for the possible selection bias.

### 5.1 Modelling selection bias for impact assessment using observables features

Since some information on the universe is available (including implementation performance), the Heckman approach can be adapted (1979)<sup>19</sup> to compute the likelihood of a project being selected for an assessment compared to the rest of the universe, conditional on the available observed characteristics. The meta-analysis is then run once again after reweighting each project by its probability of being selected into an IA.

The success of this approach, of course, rests on the number and on the importance of the observed factors in driving the selection into an IA. If one can plausibly argue that selection of IA projects depends mostly on the observable (rather than unobservable) characteristics that are observed, such as project type, region, financing and implementation performance ratings, the meta-analysis results can be adjusted for selection bias based on observables.

#### 5.1 Publication bias

The second approach considered draws on the meta-analysis literature and treat our sampling issue as a classical publication bias problem. The latter refers to the distortion of meta-analysis outcomes due to the higher likelihood of publication of statistically significant studies rather than non-significant studies. This is similar to the problem at hand – where impact assessment estimates are only available for the projects evaluated, hence less performing projects are not observed – hypothetically – in the sample. Therefore, presenting this kind of sensitivity analysis would be useful here too.

In order to test for the presence or absence of publication bias, first, a funnel plot can be used. In essence, studies are plotted on a scatter plot with effect size on the x-axis and precision or total sample size on the y-axis. If the points form an upside-down funnel shape, with a broad base that narrows towards the top of the plot, this indicates the absence of a publication bias. On the other hand, if the plot shows an asymmetric shape, with no points on one side of the graph, then publication bias can be suspected. Second, to test publication bias statistically, Begg and Mazumdar's rank correlation test or Egger's test can be used. If publication bias is detected, the trim-and-fill method can be used to correct the bias (Shi et al, 2019).

A known limitation of Trim-and-Fill is that it can correct for publication bias that does not exist, underestimating effect sizes (Terrin et al, 2003). Results of this method are therefore optional and presented in Annex IV.

---

<sup>19</sup> Heckman, J. (1979). Sample Selection Bias as a Specification Error. *Econometrica*, 47(1), 153-161.

## **6 Results from the Sensitivity Analyses**

### **6.1 Subgroup meta-analyses**

Subgroup meta-analyses are presented to assess whether the magnitude of the impact is associated with the rating class to which a project belongs, and verify whether there is a gradient between impact magnitude and rating scale. The rationale for this is due to the fact that in Section 4 it was shown how two baseline ratings were in fact unbalanced between the IFAD10 IA sample of projects and the rest of the universe.

Specifically, two features (Acceptable Disbursement Rate and Counterparts Funds) were significantly different at 5% level between projects evaluated and the unselected ones in the universe. Therefore, the extent of bias is investigated, notably whether there is a relationship between positive ratings and impacts. If bias existed, one would expect to see patterns or gradients such as the following, namely the higher the rating the higher the impact in the IFAD10 estimates of the aggregate effect sizes, as this is the main argument by IOE, notably that impact is overestimated due to higher performing projects.

Results across the Strategic Objectives (SOs), notably Market access, Resilience, Production as well as the cross-cutting theme Nutrition and the overarching IFAD goal of Economic mobility, are presented in the Annex (). Forest plots, the visual representations commonly used for meta-analytic results are employed for the purpose. The reader needs to focus on the diamonds, which represent the size of the effect – grouped according to ratings: unsatisfactory (below 3), moderately satisfactory (4), and satisfactory plus (ratings higher than 5). The “overall” diamond represents our pooled effect size, the one presented to the Board in September.

It is remarkable to see how there is no clear relationship between the ratings class and the impact estimates, particularly in the case of disbursement performance ratings. This is reassuring and corroborates the absence of bias in the impact estimates due to selection of higher performing projects.

For example, note how in the case of the market access SO domain, projects rated unsatisfactory in terms of disbursement performance ratings, display the highest impact (1.89 equivalent to 89%), compared with satisfactory plus (57%) and moderately satisfactory ones (80%). This is also largely true for the other SOs.

Also in the case of the performance rating “counterpart funds”, note how projects rated as unsatisfactory have a higher impact (29 %) compared with the moderately satisfactory ones (18%), in the domain of Production. Tables are presented in Annex II.

Therefore, a major conclusion that can be drawn from this analysis is that the direction of the ratings e.g. stronger performance, does not correspond to a larger impact.

### **6.2 Sample selection bias correction á la Heckman**

In this section, one of the sensitivity analyses described in Section 6, palatable for adjusting for the bias, is presented. The second method – trim and fill – adjusts for bias non-parametrically and given its low reliability, is presented in the Annex.

In the approach presented here, a three-stage estimation procedure is applied, whereby the results are corrected for the presence of observable and unobservable selection using an approach a la Heckman (1979), combined with meta-regression and meta-analysis.

In the first stage, a probit regression model is estimated over the universe of projects on the variables determining the selection, notably the ones that are assumed as observable drivers for the selection into IA (project sector, region, and the significant performance ratings). These estimates are used to obtain an Inverse Mills Ratio (IMR), and introduce the latter in a meta-regression in order to estimate corrected standard errors. The variance associated with this corrected standard error is then computed and summed to the original variance, to obtain final "corrected standard errors" (second stage). These final corrected standard errors are then meta-analyzed along with the original effect sizes to derive a pooled effect size adjusted for bias (third stage).

It is important to note that IMR is only marginally significant (at 10% level) in the meta-regression pertaining to the market access SO domain. For the rest of SOs it is not significant - this implies that the hypothesis of selection bias is rejected in the case of these SOs, and that it weakly holds in the context of market access.

Results are summarized in Table 6, where "observed" refers to the original impact estimates presented in the official IFAD10 Report and "adjusted" refers to the ones corrected for sample selection bias. The full set of tables is presented in the Annex III.

Note how the results maintain the integrity of the baseline random-effect meta-analysis model - the one presented in the original IFAD10 Report - and show minimal discrepancies. Specifically, results based on this scenario also remain largely positive, and indicate that, ceteris paribus, impact is overestimated by 2% in case of economic mobility, 15% in case of market access, 10% in case of production, and 6% in the case of resilience. Nutrition results remain unchanged.

However, given the lack of significance of the IMR in the second stage regression, Management concludes that there is no selection bias in the corporate impact estimates and that bias adjusted estimates are not needed and represent an over-correction.

**Table 6: Results from the Sample selection bias correction**

<b>Production</b>				
	N. projects	ES	Lower CI	Upper CI
<b>Observed</b>	17	1.44	1.26	1.64
<b>Adjusted</b>	17	1.33	1.24	1.43
<b>Market Access</b>				
	N. projects	ES	Lower CI	Upper CI
<b>Observed</b>	16	1.76	1.45	2.14
<b>Adjusted</b>	16	1.51	1.32	1.72
<b>Resilience</b>				
	N. projects	ES	Lower CI	Upper CI
<b>Observed</b>	17	1.13	1.02	1.25
<b>Adjusted</b>	17	1.06	1.02	1.1
<b>Nutrition</b>				
	N. projects	ES	Lower CI	Upper CI
<b>Observed</b>	16	1.01	0.99	1.03

<b>Adjusted</b>	16	1.01	0.99	1.02
<b>Economic Mobility</b>				
	N. projects	ES	Lower CI	Upper CI
<b>Observed</b>	17	1.74	1.51	1.97
<b>Adjusted</b>	17	1.72	1.42	2.02

## 7 Conclusions on IFAD10

In this document, a number of sensitivity analyses are presented, to assess the presence, direction and magnitude of the possible selection bias inherent in the sampling of projects chosen to be evaluated under IFAD10.

Results highlight that the bias is absent and, if anything, negligible. Through a detailed descriptive analyses it is shown that almost all the pre-determined e.g. baseline features of the IFAD10 IA sample and of the ones of the unselected projects are largely balanced – e.g. they are not statistically different - with the only exception of a couple of implementation ratings. Upon further investigation, it was found that the direction of the ratings does not imply a larger estimated impact, allowing one to conclude that projects rated highly unsatisfactory on certain attributes exhibit higher effect sizes compared with satisfactory projects. This finding strongly hints that corrective actions are put in place by implementers across the project lifetime to influence ratings towards more positive ones, particularly at completion. This factor corroborates Management's choice of not employing ratings at completion (PCR ratings) for an assessment of selection bias as the latter are endogenous (e.g., influenced by the evaluation process) and may be inflated by many reasons, the first being that ratings may reflect corrective actions by implementers, and second, that ratings do incorporate the findings of the impact assessments when available.

This analysis is complemented by an assessment of the need to correct for sample selection bias. To this end, two approaches are considered, notably the sample selection bias correction a la Heckman and the trim-and-fill approach. These are meant to more formally assess the presence and the magnitude, respectively, of any possible sample selection bias.

Given that information about the observable factors that might influence selection are available in the system, a sample selection bias correction a la Heckman is the preferred approach and is applied in the meta-analytic context.

Results based on this scenario show that selection bias does not hold and it is weakly present only in the estimations of corporate impact for market access. After computing bias adjusted estimates – the latter remain largely positive, and indicate that, *ceteris paribus*, impact was overestimated by 2% in case of economic mobility, 15% in case of market access, 10% in case of production, and 6% in the case of resilience. Nutrition results remain unchanged.

The trim-and-fill method is instead a popular tool to detect and adjust for publication bias, in other words the bias originated by ad-hoc inclusion of projects/studies in the meta-analysis. However this approach is strongly criticized by the literature (Terrin 2003, Simonsonh et al 2014) whereby meta-analysts are not recommended to perform the trim-and-fill method when using meta-analysis software programs (Shi et al, 201920), as outliers and the pre-specified direction of missing studies could have influential impact on the trim-and-fill results. In addition a known limitation of Trim-and-Fill is that it can correct for publication bias that does not exist, underestimating effect sizes (Terrin et al 2003).

Although results adjusted using this approach remain largely positive they are presented in the Annex for the above mentioned considerations.

---

<sup>20</sup> Ref : need to put all references in footnote.



## 8 Implications for IFAD11

Turning to IFAD11 – what are the implications moving forward? In the following sections the process for selecting the IFAD11 IA is summarized, and descriptive analyses are presented, to assess for the presence of selection at the time of the projects' choice in the IFAD11 context, using performance ratings and other features of the portfolio universe.

Management has formalized the process of identifying candidate IFAD-supported projects to undergo ex post impact assessments. All regional divisions have been requested to identify and select potential countries and projects to conduct impact assessments from a list of all projects scheduled to close during IFAD11 (between 2019 and 2021) as of July 2018. Projects have been identified and selected through a participatory approach which involved Management and specifically the Research and Impact Assessment Division (RIA) and each of the five regional divisions, similar to the one implemented during IFAD10.

A first screening was done in July 2018 based on disbursement rate, timing of the project, and type of project. After this first screening, further identification was conducted based on learning potential, feasibility of conducting impact assessment given the eligibility and targeting criteria and project implementation, quality of M&E data, number of beneficiaries, type of interventions, and buy-in from country and project teams. RIA staff met with representatives with each regional division to select IFAD11 ex-post impact assessments.

During this meeting, each regional division received a list of projects that RIA staff had pre-screened<sup>21</sup>. RIA staff requested each regional division to identify six projects as candidates for impact assessments during IFAD11 (two projects per replenishment year).

Subsequently, a validation exercise was conducted through follow-up meetings in collaboration with each regional division and projects received clearance from both Country and Regional Directors. Additionally, RIA held internal discussions to ensure that projects selected were representative of the IFAD11 portfolio in terms of both regional distribution and sector.

### 8.1 IFAD11 Sample Validation

As part of the IFAD11 impact assessment agenda, 24 out of 121 projects have been selected for rigorous impact assessment equalling 19.8% of total projects, 20.7% of total financing, and 23.3% of total IFAD financing. Of the 121 projects belonging to the IFAD11 universe, nine<sup>22</sup> were projects already part of evaluations initiated during IFAD9 and IFAD10 whose closing dates now fall during IFAD11. This gives a final universe of 112 projects eligible for evaluation in IFAD11. Considering the latter, the projects selected to be evaluated during IFAD11 account for 21.4% of the portfolio, representing 20.9% of total financing and 25.6% of IFAD financing.

<sup>21</sup> The number of pre-screened projects scheduled to close between 2019 and 2021 that the RIA team had initially offered to each regional division were as follows: 26 projects for APR, 21 projects for ESA, 23 projects for LAC, 17 projects for NEN, and 16 projects for WCA.

<sup>22</sup> The universe of "121 projects" include all projects CLOSING during IFAD11. However, upon further scrutiny it appeared that in the UNIVERSE of 121 – there are 9 projects whose evaluations were carried out during IFAD9 and IFAD10. The IFAD9 & IFAD10 projects are those whose closing dates were extended into IFAD11. The projects evaluated during IFAD10 are: Sao Tome and Principe PAPAC 1100001687; Senegal PAFA (extended as PAFA-E) 1100001693; Rwanda PRICE 1100001550; Kenya SDCP 1100001305; Nepal HVAP 1100001471; Bangladesh CCRIP 1100001647. The projects evaluated in IFAD9 then extended are : Uganda VODP2 1100001468; Ghana GASIP 1100001678; Bangladesh PACE 1100001648.

Table 7 presents the IFAD11 projects selected for impact assessment and their distribution by region, country, and project sector or type.

**Table 7: IFAD11 Impact Assessments by Region, Country, Project Sector and Name**

	<i>Region</i>	<i>country</i>	<i>sector</i>	<i>Project name</i>		<i>REGION</i>	<i>country</i>	<i>sector</i>	<i>Project name</i>
<b>1</b>	APR	India	CREDI	PT-Tamil Nadu	<b>13</b>	LAC	Argentina	MRKTG	PRODERI
<b>2</b>	APR	Pakistan	RURAL	SPPAP - PK	<b>14</b>	LAC	Bolivia	RURAL	ACCESOS
<b>3</b>	APR	Papua New Guinea	AGRIC	PPAP	<b>15</b>	LAC	Peru	RSRCH	PSSA
<b>4</b>	APR	Philippines	FISH	FishCORAL	<b>16</b>	NEN	Djibouti	RURAL	PRAREV-PECHE
<b>5</b>	APR	Solomon Islands	RURAL	RDP II	<b>17</b>	NEN	Kyrgyzstan	LIVST	LMDP
<b>6</b>	APR	Viet Nam	RURAL	AMD	<b>18</b>	NEN	Moldova, Republic of	RURAL	IRECR
<b>7</b>	ESA	Kenya	AGRIC	UTaNRMP	<b>19</b>	NEN	Morocco	AGRIC	PDFAZMH
<b>8</b>	ESA	Lesotho	RURAL	SADP	<b>20</b>	NEN	Tunisia	AGRIC	PRODESUD II
<b>9</b>	ESA	Malawi	RSRCH	SAPP	<b>21</b>	WCA	Ghana	CREDI	REP
<b>10</b>	ESA	Mozambique	AGRIC	PROSUL	<b>22</b>	WCA	Mali	CREDI	Rural Microfinance Programme
<b>11</b>	ESA	Tanzania, Un. Rep. of	MRKTG	MIVARF	<b>23</b>	WCA	Mauritania	RURAL	PASK II
<b>12</b>	ESA	Zambia	RSRCH	S3P	<b>24</b>	WCA	Nigeria	AGRIC	VCDP

Table 8 compares the regional distribution of these projects selected for impact assessment to the regional distribution of projects in the universe. Specifically, six projects were selected for impact assessment in both APR and ESA, five in NEN, four in WCA, and three in LAC. In the universe of 112 projects, there are 23 projects in LAC and NEN respectively, 32 in APR, 20 in ESA, and 17 in WCA.

**Table 8: Distribution of projects in the universe and in the IA sample by Region**

<b>UNIVERSE BY REGION</b>			<b>IAS BY REGION</b>		
<b>REGION</b>	<b>Projects</b>	<b>%</b>	<b>REGION</b>	<b>Projects</b>	<b>%</b>
<b>APR</b>	29	25.89	<b>APR</b>	6	25.00
<b>ESA</b>	20	17.86	<b>ESA</b>	6	25.00
<b>LAC</b>	23	20.54	<b>LAC</b>	3	12.50
<b>NEN</b>	23	20.54	<b>NEN</b>	5	20.83
<b>WCA</b>	17	15.18	<b>WCA</b>	4	16.67
<b>TOTAL</b>	112	100	<b>Total</b>	24	100.00

Table 9, first column, presents the current regional distribution of impact assessments in the IFAD11 sample and compares this distribution with three others, one weighted by the actual proportion of projects in each region and two more weighted by the actual proportional allocation of projects by financing and IFAD financing in the portfolio. As

shown in the table, the distribution of projects dictated by these different weighting schemes differ slightly from the distribution of those actually selected.

**Table 9: Distribution of IFAD11 IAs. Current and proportional to project numbers and financing (total and IFAD only).**

	<b>Actual IFAD11 IA Distribution</b>	<b>Distribution by # of Projects</b>	<b>Distribution by Financing</b>	<b>Distribution by IFAD Financing</b>
<b>APR</b>	6	6	8	7
<b>ESA</b>	6	4	6	5
<b>LAC</b>	3	5	3	3
<b>NEN</b>	5	5	3	4
<b>WCA</b>	4	4	4	5
	24	24	24	24

IFAD projects are also classified into eight project sectors (or types) in the universe (Table 10). IFAD11 projects are concentrated in rural development (46), agricultural development (29), and credit provision (16). The number of projects in the remaining five sectors range from two to six with the lowest concentration of projects in fisheries. The projects selected for IAs are comparatively less concentrated; eight rural development projects, six agricultural development projects, and three credit projects were selected. Despite the large number of credit projects in the portfolio, an equal number of projects (3) was selected in the research category along with two market access projects and one each in fisheries and livestock, respectively. No irrigation projects were selected for impact assessment during IFAD11.

**Table 10: Distribution of projects in the universe and in the IA sample by Project Sector**

<b>Universe by Project Sector</b>		<b>Project IAs sample by Project Sector</b>	
<b>Sector</b>	<b>Projects %</b>	<b>Sector</b>	<b>Projects %</b>
<b>AGRIC</b>	29 25.89	<b>AGRIC</b>	6 25
<b>CREDI</b>	16 14.29	<b>CREDI</b>	3 12.5
<b>FISH</b>	2 1.79	<b>FISH</b>	1 4.17
<b>IRRIG</b>	4 3.57	<b>IRRIG</b>	0 0
<b>LIVST</b>	6 5.36	<b>LIVST</b>	1 4.17
<b>MRKTG</b>	5 4.46	<b>MRKTG</b>	2 8.33
<b>RSRCH</b>	4 3.57	<b>RSRCH</b>	3 12.5
<b>RURAL</b>	46 41.07	<b>RURAL</b>	8 33.33
<b>Total</b>	112 100	<b>Total</b>	24 100

Looking at Table 11, note how the distribution of the IFAD11 IAs sample by sector differs from what would be dictated by the proportion of projects by sector in the universe, as well as the sectoral distribution proportional to the amount of total financing and IFAD financing. Certainly, when considering the projects by their sectoral classifications, irrigation projects seem to be underrepresented in the IFAD11 IA selection.

**Table 11: Distribution of IFAD11 IAs by project sector. Current and proportional to project numbers and financing (total and IFAD only) by sector.**

	<b>IFAD11 IAs Sample Distribution</b>	<b>Distribution by # of Projects</b>	<b>Distribution by Financing</b>	<b>Distribution Total by Financing</b>
<b>AGRIC</b>	6	6	6	7
<b>CREDI</b>	3	3	4	4
<b>FISH</b>	1	1	0	0
<b>IRRIG</b>	0	1	1	1
<b>LIVST</b>	1	1	1	1
<b>MRKTG</b>	2	1	1	1
<b>RSRCH</b>	3	1	2	1
<b>RURAL</b>	8	10	9	9
<b>Total</b>	<b>24</b>	<b>24</b>	<b>24</b>	<b>24</b>

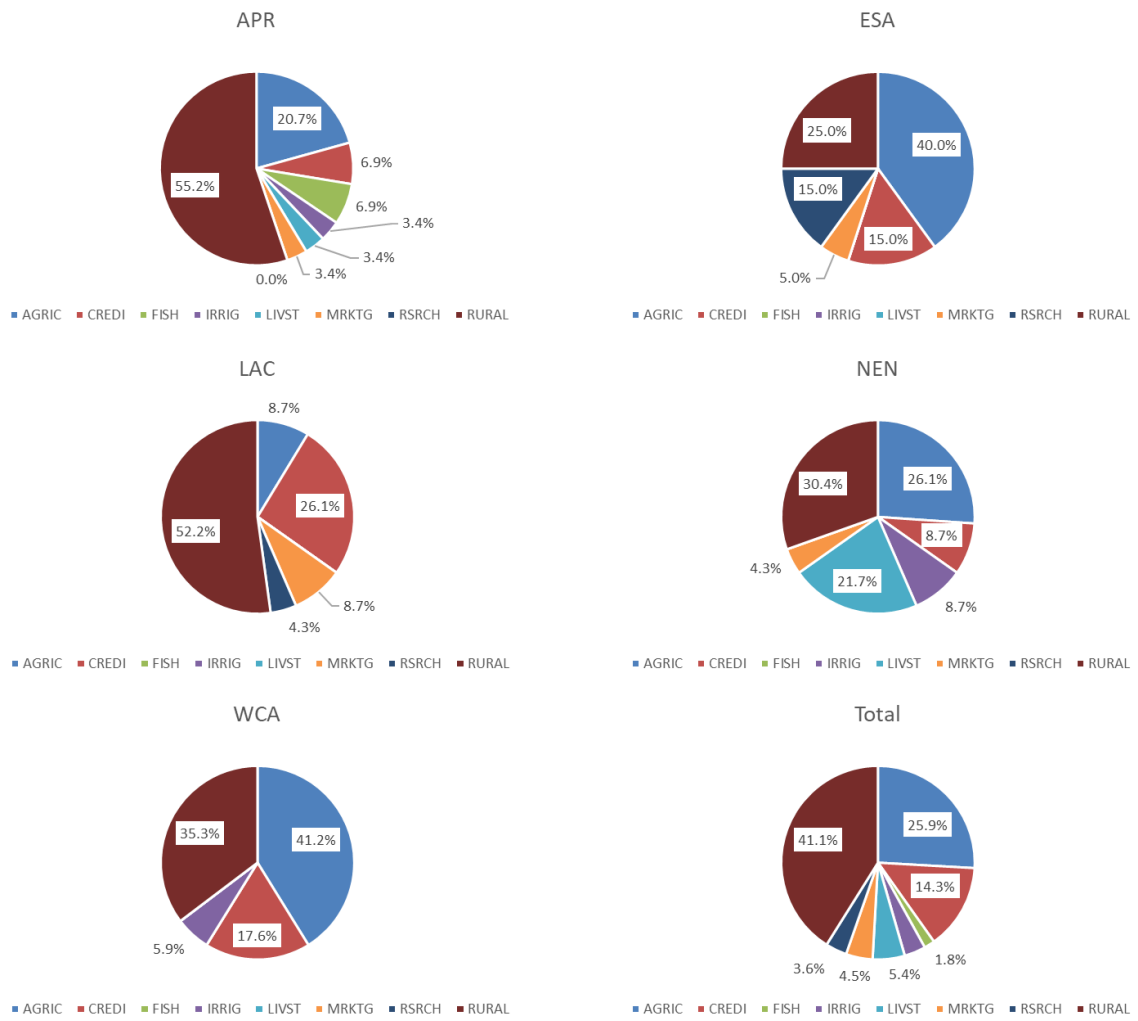
Table 12 presents descriptive statistics of the sample and the unselected projects summarizing projects by their number of beneficiaries, financing, components, and implementation score. It also presents the results of t-tests for difference in means between the sample and unselected projects to assess for the possible presence of selection bias across the various samples.

**Table 12: Balance test: selected indicators for IFAD11 (baseline).**

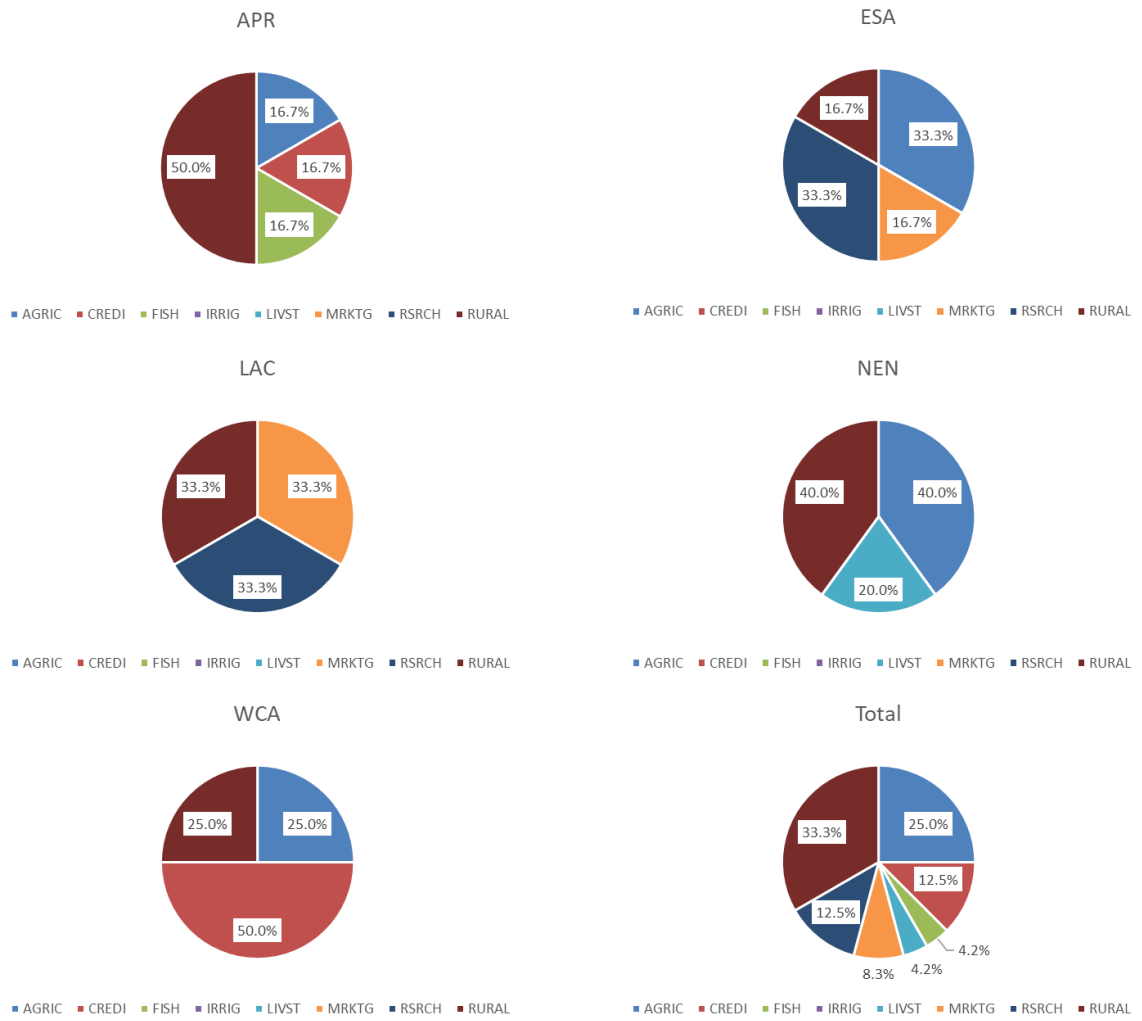
	<b>IAs Sample</b>	<b>Unselected Projects</b>	<b>Sample - Unselected</b>	
	<b>Mean</b>	<b>Mean</b>	<b>Diff. in Means</b>	<b>p-score</b>
<b>Beneficiaries</b>	476,109	1,033,194	-557,085	0.595
<b>Financing</b>	81,820,640	74,454,692	7,365,947	0.695
<b>IFAD Financing</b>	37,848,004	29,311,323	8,536,680	0.176
<b># of Financiers</b>	2.67	2.59	0.08	0.479
<b># of Subcomponent</b>	5.92	6.10	-0.18	0.753

On average, the sample of IFAD11 IAs has 476,109 beneficiaries, \$81.8 million in financing, \$37.8 million in IFAD financing, 2.6 types of financiers, and 6 subcomponents. Compared to the average across the universe of IFAD11 projects, there are 942,862 beneficiaries, \$78.7 million in financing, \$31.9 million in IFAD financing, 2.6 types of financiers, and 6 subcomponents. Note how there are no statistically significant differences across the variables presented in Table 12, across the universe and the unselected projects.

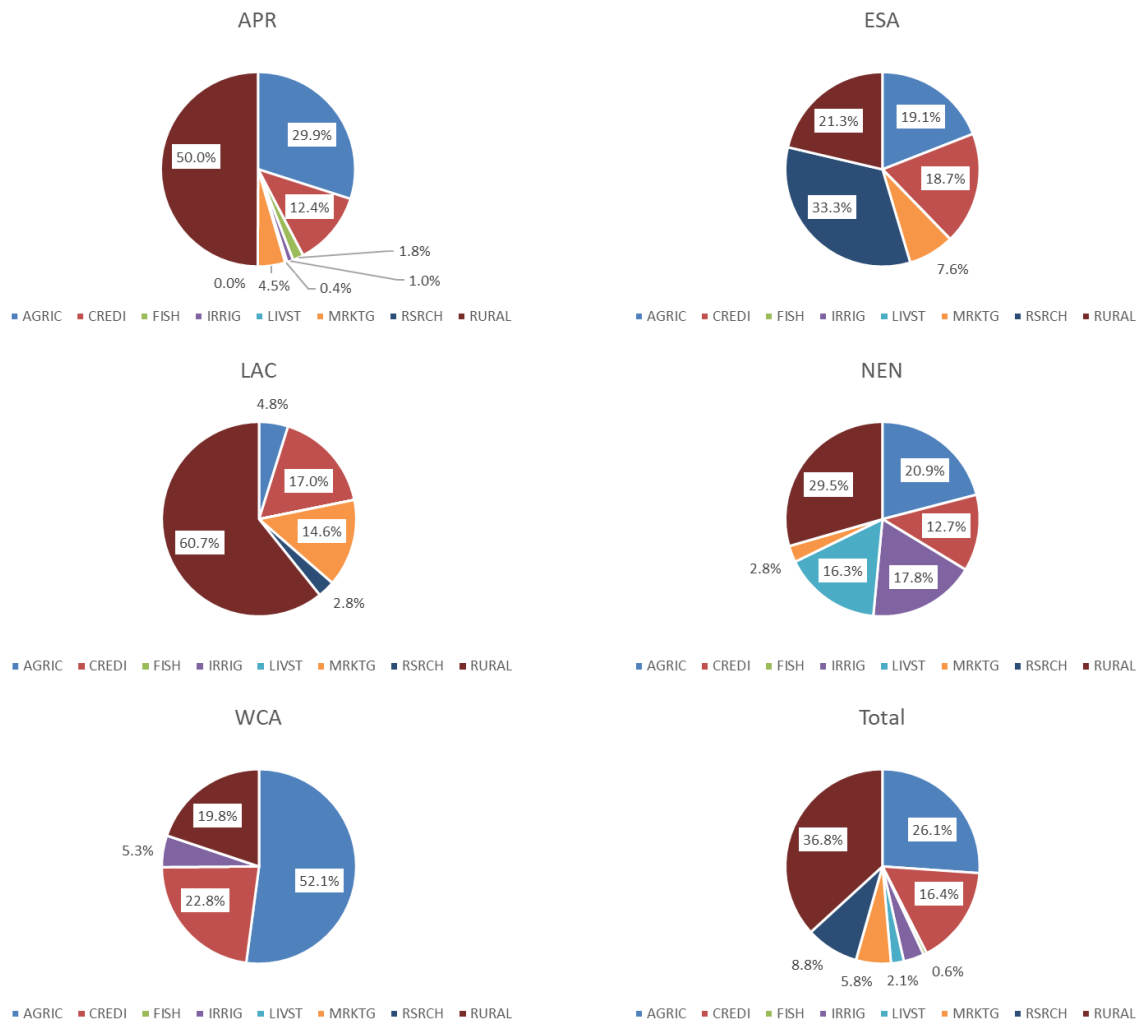
**Table 13: Distribution of IFAD11 Universe by Project Type and Region**



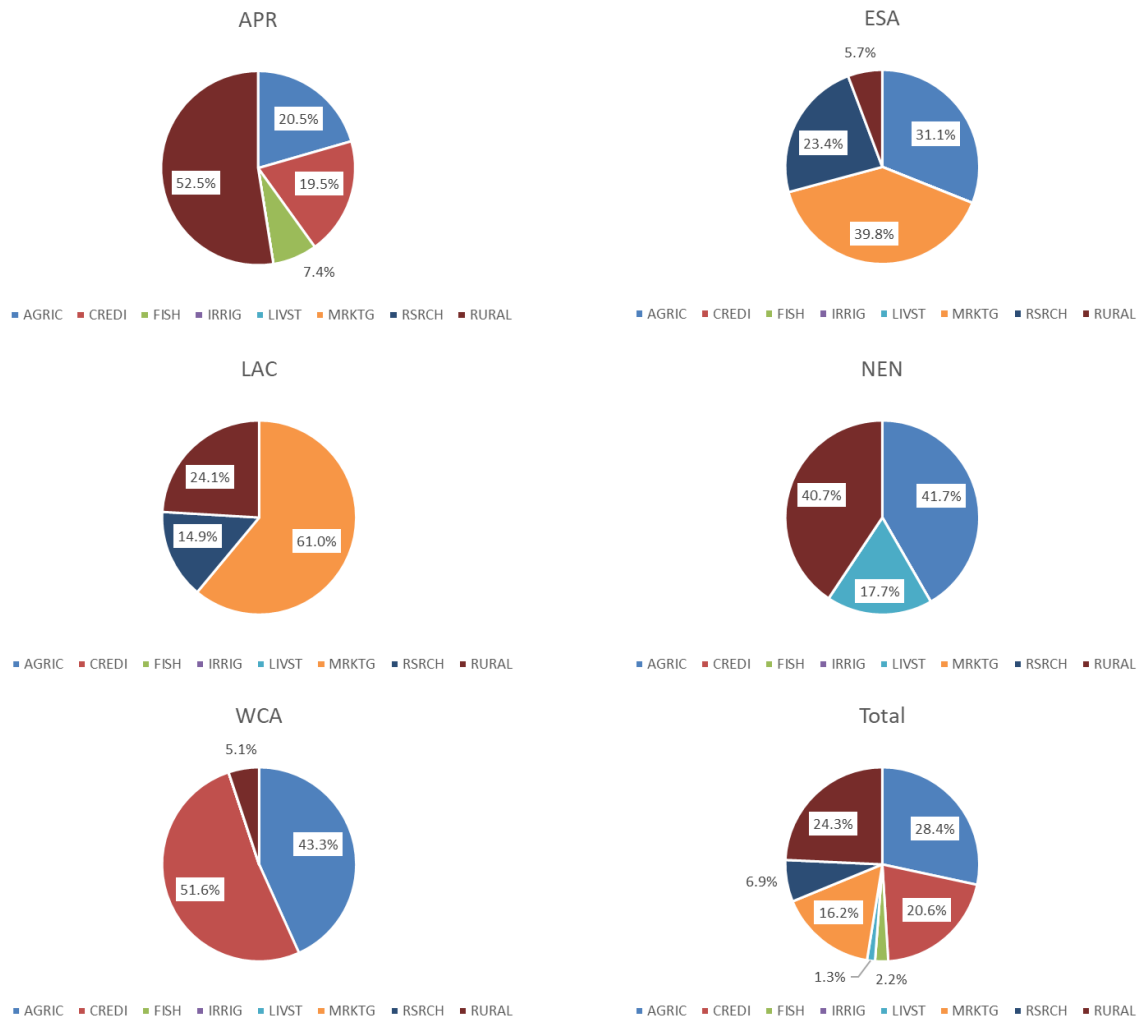
**Table 14: Distribution of IFAD11 IA Sample by Project Type and Region**



**Table 15: Distribution of IFAD11 Universe by Project Type and Region Financing**

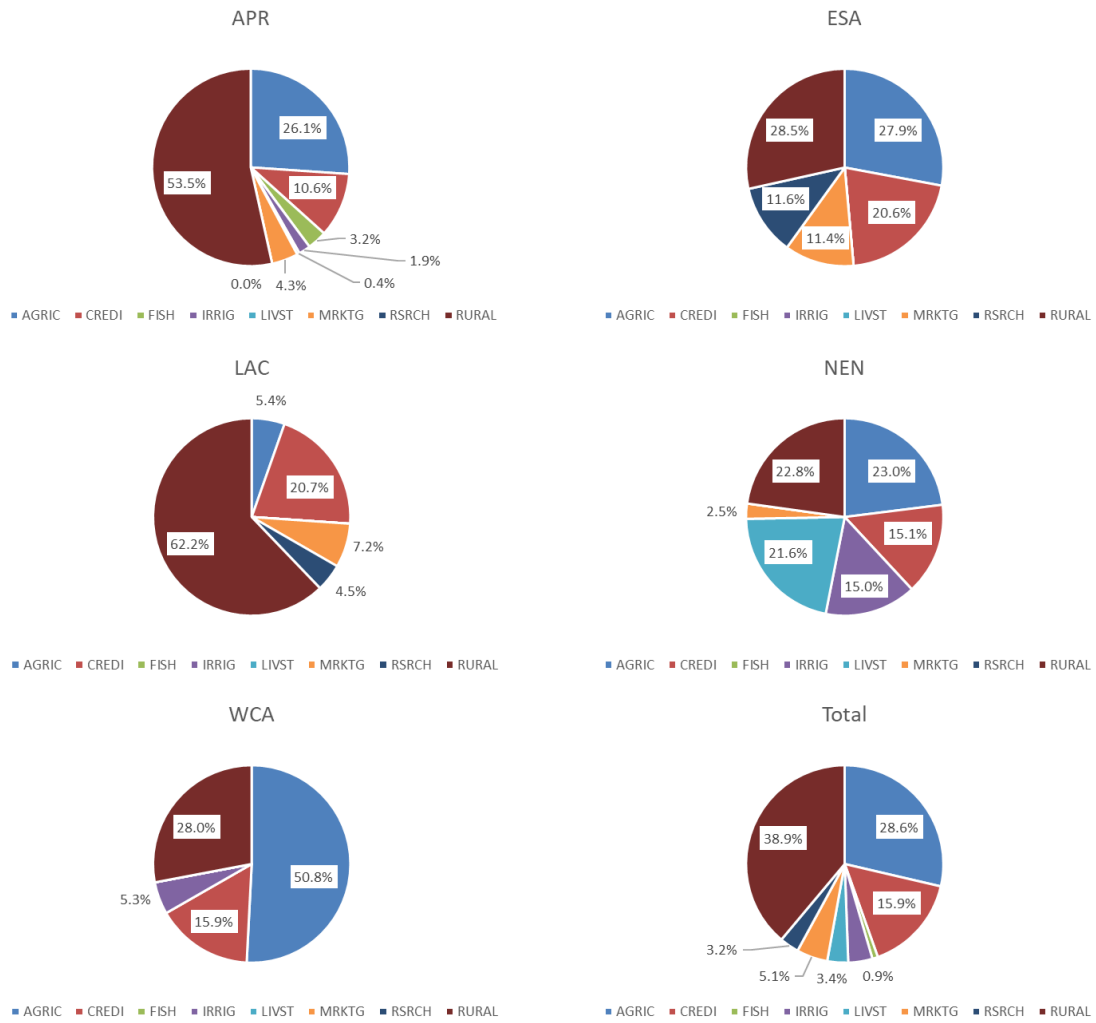


**Table 16: Distribution of IFAD11 IA Sample by Project Type and Region Financing**

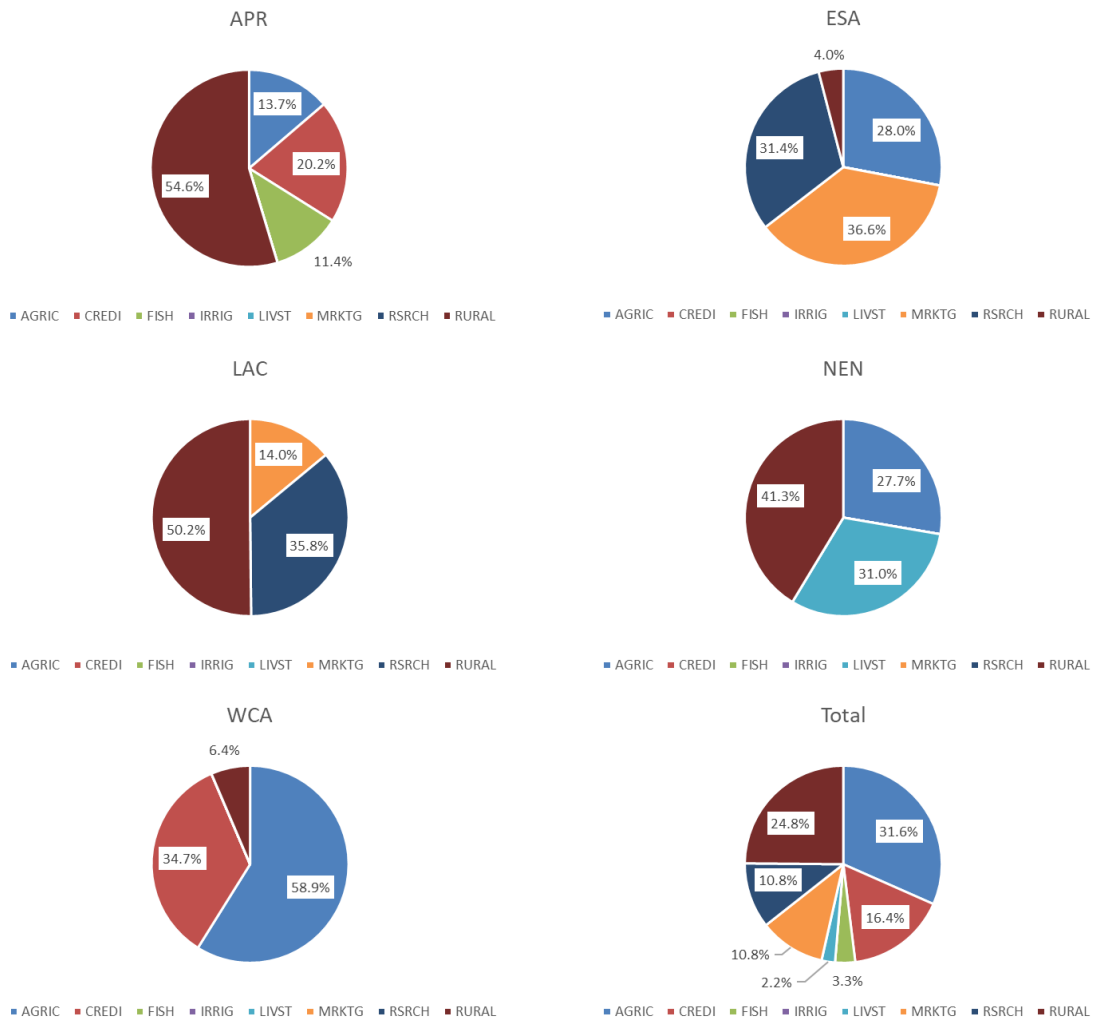




**Table 17: Distribution of IFAD11 Universe by Project Type and Region IFAD Financing**



**Table 18: Distribution of IFAD11 IA Sample by Project Type and Region IFAD Financing**



In addition to being divided into project sector and regional categories, projects can further be disaggregated into component, subcomponent and subcomponent type. The only variable that is standardized in the system (GRIPS) is the subcomponent type, which can be broken down into approximately 60 categories. Note that the component feature is not standardized across IFAD databases and therefore cannot be used in this analysis.

Each project contains multiple subcomponents funded by different sources. Nearly all projects contain a management and coordination component, but there remains considerable heterogeneity. Apart from the management subcomponent, the most common subcomponent in the universe of projects is rural financial services (44.7%), followed by local capacity building (38.39%), monitoring and evaluation (28.57%), and institutional support (27.68%). Table 19 compares the relative distribution of subcomponent types in the selected IAs and the IFAD11 portfolio.

When projects are categorized by their largest subcomponent (in terms of current financing), the most common subcomponents among IFAD11 IAs are development funds and rural financial services each with three projects selected for IAs, respectively (table available upon request). This matches the portfolio overall where development funds and rural financial services are the largest subcomponents with thirteen and fourteen projects, respectively. However, beyond that similarity, the distribution of IAs unevenly represents the distribution in the portfolio. Finally, looking at Table 20– the distribution of projects by largest financier and region is provided in the portfolio and in the IA sample.

Two issues in this analysis hinder both our ability to draw a representative sample by sector and components distribution and assess the representativeness of our current selection. The first is that the sector variable does not reflect the true nature of the project. For example, projects in which the true intervention is related to livestock or animal husbandry may be classified as marketing if there is substantial intervention in business formation or a rural development project may have a substantive irrigation component but not be classified as such because it is combined with other interventions. Moreover, the sectors are too broad such that there is substantial heterogeneity within sectors, namely the rural and agricultural development. Secondly, the subcomponent type is too disaggregated (60 unique entries over 112 records) to be used as a possible stratification feature hence it would need to be recoded prior to be used. However, there currently exists no method for standardizing or harmonizing subcomponents within projects and each project can have multiple subcomponents all of varying sizes and relative importance.

**Table 19: Distribution of subcomponent types in the IAs sample and in the IFAD11 universe**

<b>Subcomponent Type</b>	<b>IAs</b>		<b>Universe</b>	
	<b>Freq.</b>	<b>%</b>	<b>Freq.</b>	<b>%</b>
<i>Rural financial services</i>	<b>10</b>	41.67	<b>50</b>	44.64
<i>Local capacity building</i>	<b>8</b>	33.33	<b>43</b>	38.39
<i>Monitoring and evaluation</i>	<b>7</b>	29.17	<b>32</b>	28.57
<i>Institutional support</i>	<b>5</b>	20.83	<b>31</b>	27.68
<i>Technology transfer</i>	<b>7</b>	29.17	<b>27</b>	24.11
<i>Rural infrastructure</i>	<b>5</b>	20.83	<b>25</b>	22.32
<i>Business development</i>	<b>2</b>	8.33	<b>21</b>	18.75
<i>Development funds</i>	<b>5</b>	20.83	<b>20</b>	17.86
<i>Irrigation infrastructure</i>	<b>1</b>	4.17	<b>19</b>	16.96
<i>Community development</i>	<b>5</b>	20.83	<b>18</b>	16.07
<i>Crop production</i>	<b>3</b>	12.50	<b>17</b>	15.18
<i>Marketing: inputs/outputs</i>	<b>6</b>	25.00	<b>17</b>	15.18

<i>Roads/tracks</i>	<b>1</b>	4.17	<b>17</b>	15.18
<i>Micro-enterprises</i>	<b>3</b>	12.50	<b>16</b>	14.29
<i>Climate change adaptation</i>	<b>4</b>	16.67	<b>14</b>	12.50
<i>Rangelands/pastures</i>	<b>2</b>	8.33	<b>14</b>	12.50
<i>Credit</i>	<b>1</b>	4.17	<b>13</b>	11.61
<i>Resource mgmt/protection</i>	<b>1</b>	4.17	<b>13</b>	11.61
<i>Rural enterprises</i>	<b>2</b>	8.33	<b>13</b>	11.61
<i>Animal husbandry</i>	<b>2</b>	8.33	<b>11</b>	9.82
<i>Technology development</i>	<b>2</b>	8.33	<b>11</b>	9.82
<i>Policy Support/Development</i>	<b>2</b>	8.33	<b>9</b>	8.04
<i>Animal health</i>	<b>1</b>	4.17	<b>8</b>	7.14
<i>Crop extension services</i>	<b>2</b>	8.33	<b>8</b>	7.14
<i>Market information/study</i>	<b>3</b>	12.50	<b>8</b>	7.14
<i>Training</i>	<b>3</b>	12.50	<b>8</b>	7.14
<i>Communication</i>	<b>1</b>	4.17	<b>7</b>	6.25
<i>Market infrastructure</i>	<b>3</b>	12.50	<b>7</b>	6.25
<i>Soil and Water conservation</i>	<b>1</b>	4.17	<b>7</b>	6.25
<i>Livestock - other</i>	<b>0</b>	0.00	<b>6</b>	5.36
<i>Livestock post-harvest</i>	<b>2</b>	8.33	<b>6</b>	5.36
<i>Seed, fertilizer, pesticide</i>	<b>1</b>	4.17	<b>6</b>	5.36
<i>Health and nutrition</i>	<b>0</b>	0.00	<b>5</b>	4.46
<i>Drinking water/sanitation</i>	<b>0</b>	0.00	<b>4</b>	3.57
<i>Forestry</i>	<b>1</b>	4.17	<b>4</b>	3.57
<i>Literacy</i>	<b>0</b>	0.00	<b>4</b>	3.57
<i>Crop technology development</i>	<b>1</b>	4.17	<b>3</b>	2.68
<i>Disaster mitigation</i>	<b>2</b>	8.33	<b>3</b>	2.68
<i>Financing/preparation charges</i>	<b>1</b>	4.17	<b>3</b>	2.68
<i>Fisheries infrastructure</i>	<b>2</b>	8.33	<b>3</b>	2.68
<i>Irrigation management</i>	<b>0</b>	0.00	<b>3</b>	2.68
<i>Land improvement</i>	<b>1</b>	4.17	<b>3</b>	2.68
<i>On-farm storage</i>	<b>1</b>	4.17	<b>3</b>	2.68
<i>Animal restocking</i>	<b>1</b>	4.17	<b>2</b>	1.79
<i>Aquaculture</i>	<b>0</b>	0.00	<b>2</b>	1.79
<i>Education (primary/second)</i>	<b>1</b>	4.17	<b>2</b>	1.79
<i>Fisheries/marine conservation</i>	<b>2</b>	8.33	<b>2</b>	1.79
<i>Input supply</i>	<b>1</b>	4.17	<b>2</b>	1.79
<i>Knowledge management</i>	<b>0</b>	0.00	<b>2</b>	1.79
<i>Land reform/titles</i>	<b>0</b>	0.00	<b>2</b>	1.79
<i>Processing</i>	<b>0</b>	0.00	<b>2</b>	1.79
<i>Standards and regulations</i>	<b>1</b>	4.17	<b>2</b>	1.79
<i>Energy production</i>	<b>0</b>	0.00	<b>1</b>	0.89
<i>Fishing (capture)</i>	<b>0</b>	0.00	<b>1</b>	0.89
<i>Housing</i>	<b>1</b>	4.17	<b>1</b>	0.89
<i>Insurance/risk transfer</i>	<b>1</b>	4.17	<b>1</b>	0.89
<i>Legal assistance</i>	<b>0</b>	0.00	<b>1</b>	0.89
<i>Mechanization services</i>	<b>0</b>	0.00	<b>1</b>	0.89
<i>Venture capital</i>	<b>1</b>	4.17	<b>1</b>	0.89

Finally, looking at Table 20 – the distribution of projects by largest financier and region is provided in the portfolio and in the IA sample.

**Table 20 Distribution of Largest Financier Type by Region.**

**IFAD11 Universe**

	<b>APR</b>	<b>ESA</b>	<b>LAC</b>	<b>NEN</b>	<b>WCA</b>	<b>Total</b>
<b>Largest Financier</b>						
<i>Domestic</i>	10	4	9	3	1	<b>27</b>
<i>IFAD</i>	13	13	12	18	14	<b>70</b>
<i>International</i>	6	3	2	2	2	<b>15</b>
<b>Total</b>	<b>29</b>	<b>20</b>	<b>23</b>	<b>23</b>	<b>17</b>	<b>112</b>

**IFAD11 IAs Sample**

	<b>APR</b>	<b>ESA</b>	<b>LAC</b>	<b>NEN</b>	<b>WCA</b>	<b>Total</b>
<b>Largest Financier</b>						
<i>Domestic</i>	1	0	1	2	1	<b>5</b>
<i>IFAD</i>	3	5	2	3	3	<b>16</b>
<i>International</i>	2	1	0	0	0	<b>3</b>
<b>Total</b>	<b>6</b>	<b>6</b>	<b>3</b>	<b>5</b>	<b>4</b>	<b>24</b>

Table **21** presents additional baseline ratings statistics based on the first available implementation performance rating available in the system - given by the corresponding supervision report. Here, only performance of M&E system seems to be statistically significant. However, given that 24 implementation ratings have been tested - these results stress the absence of selection bias in the case of the IFAD11 sample of IAs.

**Table 21: Balance tests: implementation performance ratings for IFAD11 (baseline characteristics)**

	IFAD11 IA Sample	Unselected Projects (closing during IFAD11)	IFAD 11 Universe	Sample - Unselected		Sample - Universe	
	Mean	Mean	Mean	Diff. in Means	p- score	Diff. in Means	p- score
<i>Assessment of the Overall Implementation Performance</i>	3.96	3.94	3.94	0.02	0.80	0.02	0.80
<i>Likelihood of Achieving the Development Objective Effectiveness</i>	4.04	4.02	4.04	0.02	0.82	0.01	0.94
<i>Targeting and Outreach</i>	3.91	3.96	3.95	-0.05	0.61	-0.04	0.69
<i>Gender equality &amp; women's participation</i>	4.08	4.09	4.09	0.00	0.98	-0.01	0.95
<i>Agricultural Productivity</i>	4.00	4.04	4.04	-0.04	0.71	-0.04	0.62
<i>Adaptation to Climate Change</i>	4.00	3.97	3.98	0.03	0.66	0.02	0.78
<i>Institutions and Policy Engagement</i>	4.08	4.00	4.02	0.08	0.40	0.07	0.48
<i>Human and Social Capital and Empowerment</i>	4.00	3.97	3.99	0.03	0.73	0.01	0.92
<i>Quality of Beneficiary Participation</i>	3.95	4.03	4.02	-0.07	0.39	-0.07	0.26
<i>Responsiveness of Service Providers</i>	4.04	4.05	4.05	-0.01	0.94	-0.01	0.89
<i>Environment and Natural Resource Management</i>	4.00	3.96	3.97	0.04	0.68	0.03	0.70
<i>Exit Strategy</i>	4.09	3.98	4.00	0.11	0.21	0.09	0.37
<i>Potential for Scaling-up</i>	3.93	4.00	3.99	-0.07	0.40	-0.05	0.48
<i>Quality of Project Management</i>	3.95	4.07	4.04	-0.12	0.15	-0.09	0.13
<i>Knowledge Management</i>	3.96	3.96	3.97	-0.01	0.97	-0.02	0.89
<i>Coherence between AWPB and Implementation</i>	4.00	3.99	3.99	0.01	0.87	0.01	0.89
<i>Performance of M&amp;E System</i>	3.92	3.77	3.82	0.15	0.24	0.09	0.35
<i>Acceptable Disbursement Rate</i>	4.04	3.73	3.80	0.31	0.00	0.24	0.01
<i>Quality of Financial Management</i>	2.71	3.10	3.02	-0.39	0.31	-0.31	0.44
<i>Quality and Timeliness of Audit</i>	3.96	4.00	3.98	-0.04	0.71	-0.02	0.79
<i>Counterparts Funds</i>	4.00	3.93	3.95	0.07	0.38	0.05	0.44
<i>Compliance with Loan Covenants</i>	4.08	3.99	4.01	0.10	0.50	0.07	0.64
<i>Procurement</i>	4.00	3.98	3.99	0.02	0.83	0.01	0.94
	3.83	3.94	3.92	-0.11	0.33	-0.09	0.53

## **8.2 Conclusions for IFAD11**

In light of these descriptive analyses and broad considerations, Management can conclude that there is no selection bias in the IFAD11 sample of impact assessments.

However the additional following recommendation can be made, notably adjusting the regional distribution to allow for two more impact assessments in LAC.



## Appendix – Annex I

**Table 22: Balance test: Implementation Performance Ratings (Baseline Characteristics) by region**

	APR						ESA					
	Sample		Unselected		Sample - Unselected		Sample		Unselected		Sample - Unselected	
	n	Mean	n	Mean	Difference	p-score	n	Mean	n	Mean	Difference	p-score
<b>Duration</b>	5	6.80	25	8.56	-1.76	0.31	6	10.00	14	8.29	1.71	0.07
<b>Beneficiaries</b>	5	1,226,531	25	1,108,359	118,173	0.92	6	377,717	14	464,178	-86,461	0.85
<b>Approved Funding</b>	5	74,052,700	25	62,145,347	11,907,353	0.64	6	68,786,299	14	83,084,966	-83,084,966	0.85
<b>Assessment of the Overall Implementation Performance</b>	5	4.20	25	3.88	0.32	0.07	6	4.17	14	3.71	0.45	0.17
<b>Likelihood of Achieving the Development Objective</b>	5	4.00	25	3.92	0.08	0.72	6	4.00	14	3.86	0.14	0.61
<b>Effectiveness</b>	4	3.50	19	3.84	-0.34	0.14	3	4.00	13	3.92	0.08	0.65
<b>Targeting and Outreach</b>	5	4.20	25	3.96	0.24	0.19	6	4.17	14	4.00	0.17	0.40
<b>Gender equality &amp; women's participation</b>	5	4.20	25	4.00	0.20	0.33	6	4.17	14	4.07	0.10	0.54
<b>Agricultural Productivity</b>	5	4.40	19	4.00	0.40	0.00	5	4.00	14	3.71	0.29	0.32
<b>Adaptation to Climate Change</b>	1	4.00	3	4.00	0.00	-	-	-	-	-	-	-
<b>Institutions and Policy Engagement</b>	5	4.40	21	4.00	0.40	0.10	6	4.33	14	4.00	0.33	0.36
<b>Human and Social Capital and Empowerment</b>	4	4.25	21	3.95	0.30	0.36	6	4.17	14	3.79	0.38	0.16
<b>Quality of Beneficiary Participation</b>	5	4.00	25	4.04	-0.04	0.80	6	4.00	14	4.00	0.00	-
<b>Responsiveness of Service Providers</b>	5	4.40	25	3.84	0.56	0.01	6	3.67	14	4.07	-0.41	0.11
<b>Environment and Natural Resource Management</b>	1	4.00	3	4.00	0.00	-	-	-	-	-	-	-
<b>Exit Strategy</b>	4	4.25	19	4.00	0.25	0.03	3	4.00	3	3.67	0.33	0.37
<b>Potential for Scaling-up</b>	4	4.25	19	4.00	0.25	0.03	5	4.40	14	4.14	0.26	0.51
<b>Quality of Project Management</b>	5	4.00	25	3.84	0.16	0.61	6	4.17	14	3.64	0.52	0.24
<b>Knowledge Management</b>	4	4.00	22	3.86	0.14	0.45	6	4.17	13	3.92	0.24	0.14
<b>Coherence between AWPB and Implementation</b>	5	4.00	22	3.82	0.18	0.32	6	4.17	14	3.86	0.31	0.37
<b>Performance of M&amp;E System</b>	5	3.80	24	3.88	-0.08	0.80	6	3.50	14	3.71	-0.21	0.47
<b>Acceptable Disbursement Rate</b>	5	4.20	25	3.32	0.88	0.14	6	4.17	14	3.86	0.31	0.62
<b>Quality of Financial Management</b>	5	4.40	22	3.91	0.49	0.04	5	3.80	14	3.71	0.09	0.81

<b>Quality and Timeliness of Audit</b>	5	4.40	24	3.92	0.48	0.07	6	4.00	14	3.86	0.14	0.53
<b>Counterparts Funds</b>	5	4.40	25	4.00	0.40	0.28	6	4.17	14	4.29	-0.12	0.80
<b>Compliance with Loan Covenants</b>	5	4.40	25	3.92	0.48	0.01	6	4.33	14	3.79	0.55	0.10
<b>Procurement</b>	5	4.40	25	3.96	0.44	0.03	6	3.83	14	3.79	0.05	0.88

	LAC						NEN					
	Sample		Unselected		Sample - Unselected		Sample		Unselected		Sample - Unselected	
	n	Mean	n	Mean	Difference	p-score	n	Mean	n	Mean	Difference	p-score
<b>Duration</b>	3	6.667	15	8.00	-1.33	0.35	1	7.00	12	7.92	-0.92	0.66
<b>Beneficiaries</b>	3	40,518	15	63,093	-22,576	0.62	1	145,600	12	92,023	53,577	0.58
<b>Approved Funding</b>	3	31,437,826	15	25,628,214	5,809,612	0.57	1	15,780,852	12	38,844,717	-23,063,865	0.31
<b>Assessment of the Overall Implementation Performance</b>	3	4	15	3.53	0.47	0.36	1	4.00	12	4.00	0.00	1.00
<b>Likelihood of Achieving the Development Objective</b>	3	4.00	15	3.73	0.27	0.46	1	4.00	12	4.17	-0.17	0.79
<b>Effectiveness</b>	1	4.00	10	3.60	0.40	0.66	1	4.00	8	3.75	0.25	0.75
<b>Targeting and Outreach</b>	3	4.67	15	4.13	0.53	0.20	1	4.00	12	4.42	-0.42	0.45
<b>Gender equality &amp; women's participation</b>	3	3.67	15	3.67	0.00	1.00	1	4.00	12	4.08	-0.08	0.88
<b>Agricultural Productivity</b>	1	4.00	10	3.90	0.10	0.90	1	4.00	10	3.90	0.10	0.87
<b>Adaptation to <i>Climate Change</i></b>	-	-	-	-	-	-	-	-	-	-	-	-
<b>Institutions and Policy Engagement</b>	3	3.33	11	3.82	-0.49	0.44	1	4.00	12	3.92	0.08	0.88
<b>Human and Social Capital and Empowerment</b>	1	4.00	11	3.91	0.09	0.92	1	4.00	12	3.92	0.08	0.88
<b>Quality of Beneficiary Participation</b>	3	3.67	15	3.87	-0.20	0.74	1	4.00	12	4.08	-0.08	0.88
<b>Responsiveness of Service Providers</b>	3	3.67	15	3.73	-0.07	0.83	1	4.00	12	4.08	-0.08	0.91
<b>Environment and Natural Resource Management</b>	-	-	-	-	-	-	-	-	-	-	-	-
<b>Exit Strategy</b>	1	4.00	9	3.67	0.33	0.67	1	4.00	11	4.09	-0.09	0.90
<b>Potential for Scaling-up</b>	1	4.00	10	3.80	0.20	0.81	1	4.00	11	4.27	-0.27	0.75
<b>Quality of Project Management</b>	3	4.00	15	3.67	0.33	0.60	1	3.00	12	4.08	-1.08	0.15
<b>Knowledge Management</b>	2	5.00	10	4.00	1.00	0.18	1	4.00	12	4.25	-0.25	0.71
<b>Coherence between AWPB and Implementation</b>	2	4.50	12	3.50	1.00	0.12	1	4.00	12	3.83	0.17	0.79

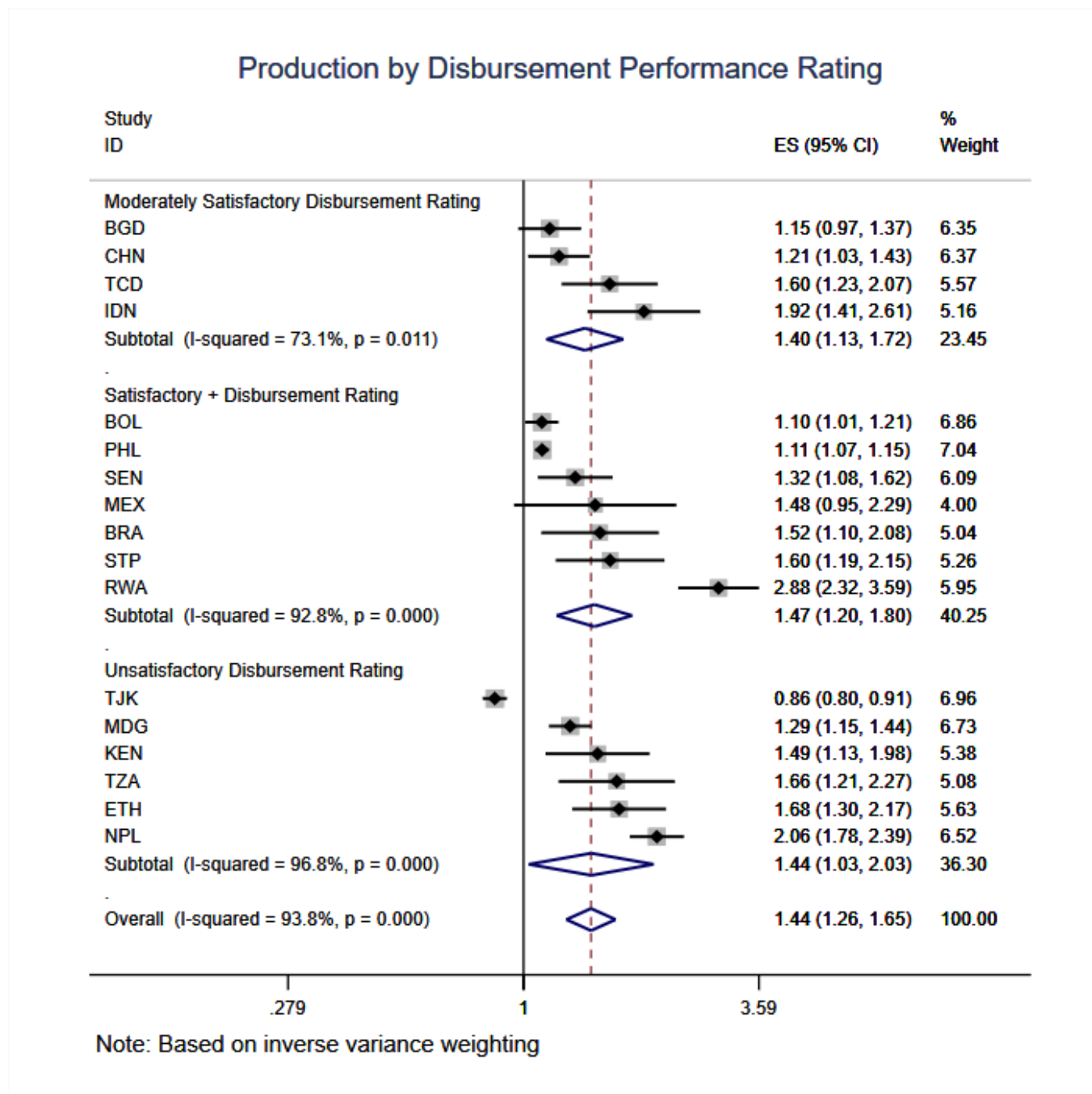
<b>Performance of M&amp;E System</b>	3	3.67	15	3.53	0.13	0.78	1	4.00	12	4.08	-0.08	0.88
<b>Acceptable Disbursement Rate</b>	3	4.33	15	3.47	0.87	0.35	1	1.00	12	3.83	-2.83	0.14
<b>Quality of Financial Management</b>	3	4.00	13	3.85	0.15	0.83	1	4.00	11	4.18	-0.18	0.68
<b>Quality and Timeliness of Audit</b>	3	4.00	15	4.00	0.00	1.00	1	4.00	12	4.25	-0.25	0.76
<b>Counterparts Funds</b>	3	4.67	15	3.67	1.00	0.06	1	4.00	12	4.33	-0.33	0.63
<b>Compliance with Loan Covenants</b>	3	4.00	15	4.00	0.00	1.00	1	3.00	12	4.17	-1.17	0.02
<b>Procurement</b>	3	4.67	15	4.07	0.60	0.28	1	4.00	12	4.08	-0.08	0.92

	WCA											
	Sample		Unselected		Sample - Unselected							
	n	Mean	n	Mean	Difference	p-score						
<b>Duration</b>	4	8.50	22	8.32	0.18	0.90						
<b>Beneficiaries</b>	4	162583	22	897,950	-735,368	0.55						
<b>Approved Funding</b>	4	22364480	22	40,751,001	-18,386,521	0.34						
<b>Assessment of the Overall Implementation Performance</b>	4	4.00	22	4.05	-0.05	0.68						
<b>Likelihood of Achieving the Development Objective</b>	4	4.00	22	4.18	-0.18	0.37						
<b>Effectiveness</b>	3	4.00	19	4.11	-0.11	0.58						
<b>Targeting and Outreach</b>	4	4.25	22	4.18	0.07	0.80						
<b>Gender equality &amp; women's participation</b>	4	4.00	22	4.09	-0.09	0.77						
<b>Agricultural Productivity</b>	3	4.00	18	4.11	-0.11	0.57						
<b>Adaptation to Climate Change</b>	1	4.00	3	4.33	-0.33	0.67						
<b>Institutions and Policy Engagement</b>	3	4.00	20	4.20	-0.20	0.52						
<b>Human and Social Capital and Empowerment</b>	3	4.00	19	4.00	0.00	1.00						
<b>Quality of Beneficiary Participation</b>	4	4.00	22	4.23	-0.23	0.41						
<b>Responsiveness of Service Providers</b>	4	3.75	22	4.14	-0.39	0.21						
<b>Environment and Natural Resource Management</b>	1	4.00	5	4.00	0.00	-						
<b>Exit Strategy</b>	2	4.00	16	4.06	-0.06	0.74						
<b>Potential for Scaling-up</b>	3	4.00	18	4.11	-0.11	0.57						
<b>Quality of Project Management</b>	4	3.75	22	4.00	-0.25	0.56						
<b>Knowledge Management</b>	3	4.00	19	4.16	-0.16	0.48						

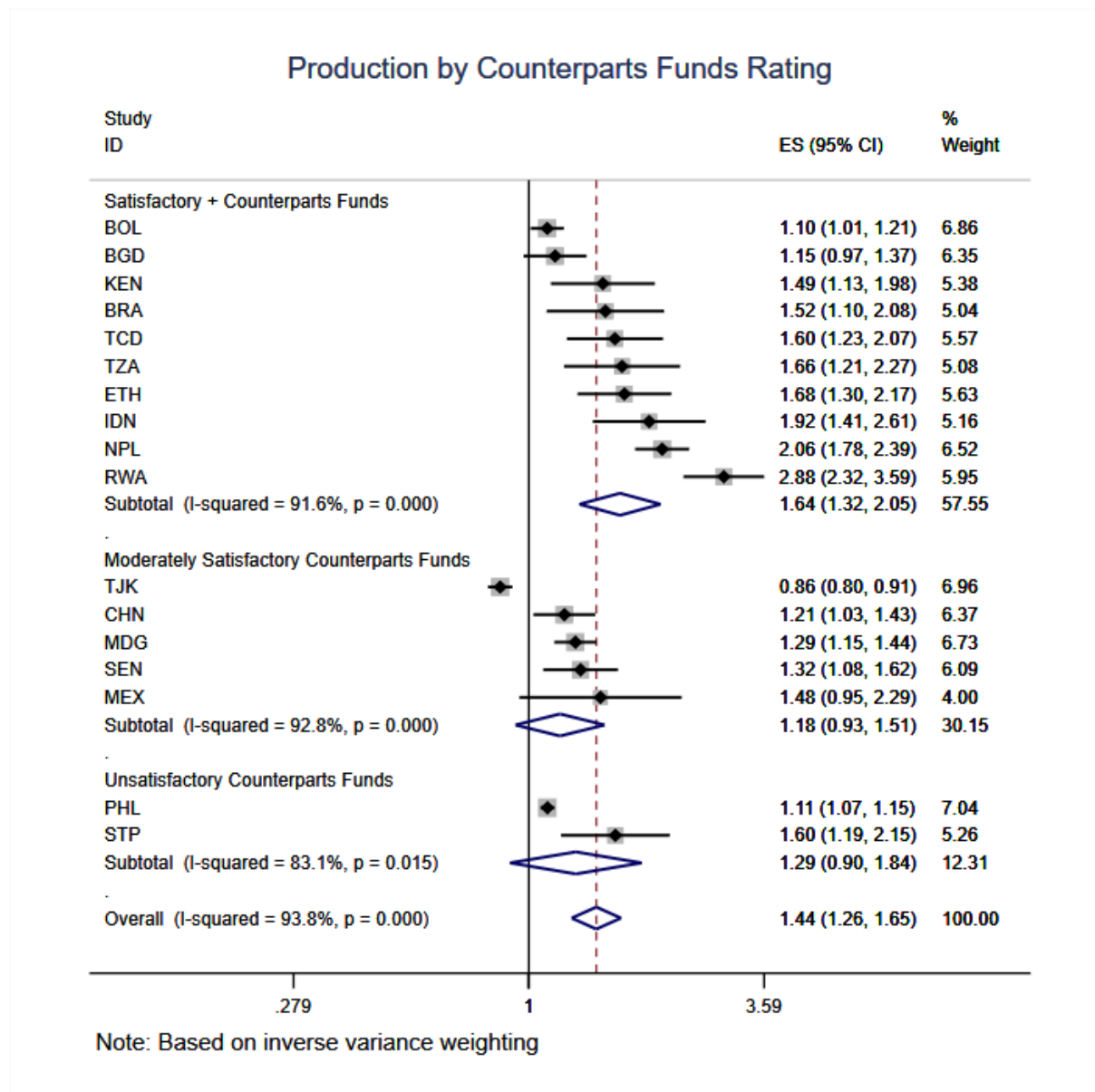
<b>Coherence between AWPB and Implementation</b>	3	4.00	21	3.86	0.14	0.79					
<b>Performance of M&amp;E System</b>	4	4.00	22	3.91	0.09	0.81					
<b>Acceptable Disbursement Rate</b>	4	5.00	22	3.05	1.96	0.02					
<b>Quality of Financial Management</b>	3	4.33	19	3.90	0.44	0.23					
<b>Quality and Timeliness of Audit</b>	4	4.00	22	4.09	-0.09	0.86					
<b>Counterparts Funds</b>	4	4.75	22	3.91	0.84	0.01					
<b>Compliance with Loan Covenants</b>	4	4.25	22	4.23	0.02	0.93					
<b>Procurement</b>	4	4.00	22	4.09	-0.09	0.68					

## Appendix - Annex II

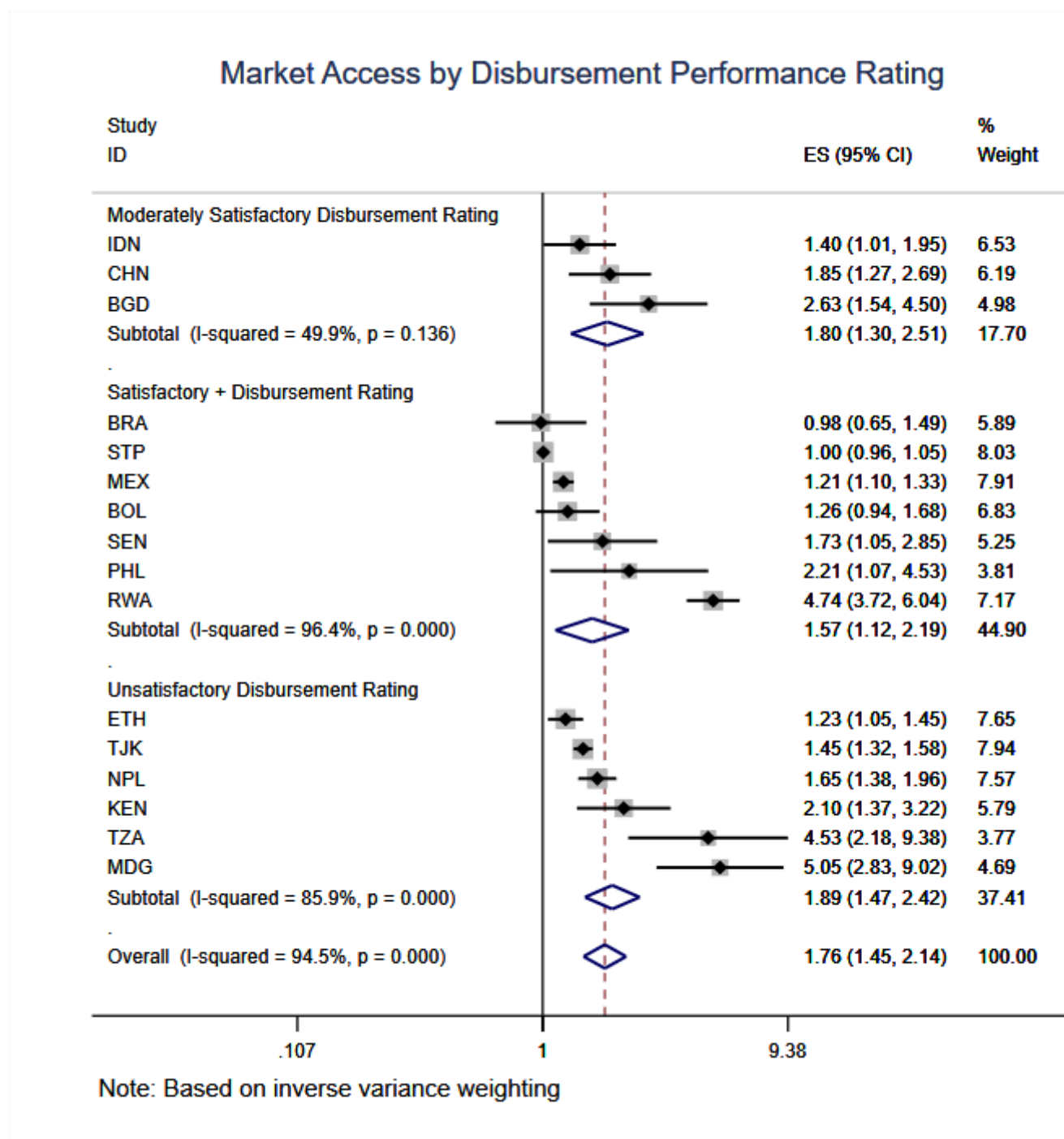
**Table 23**



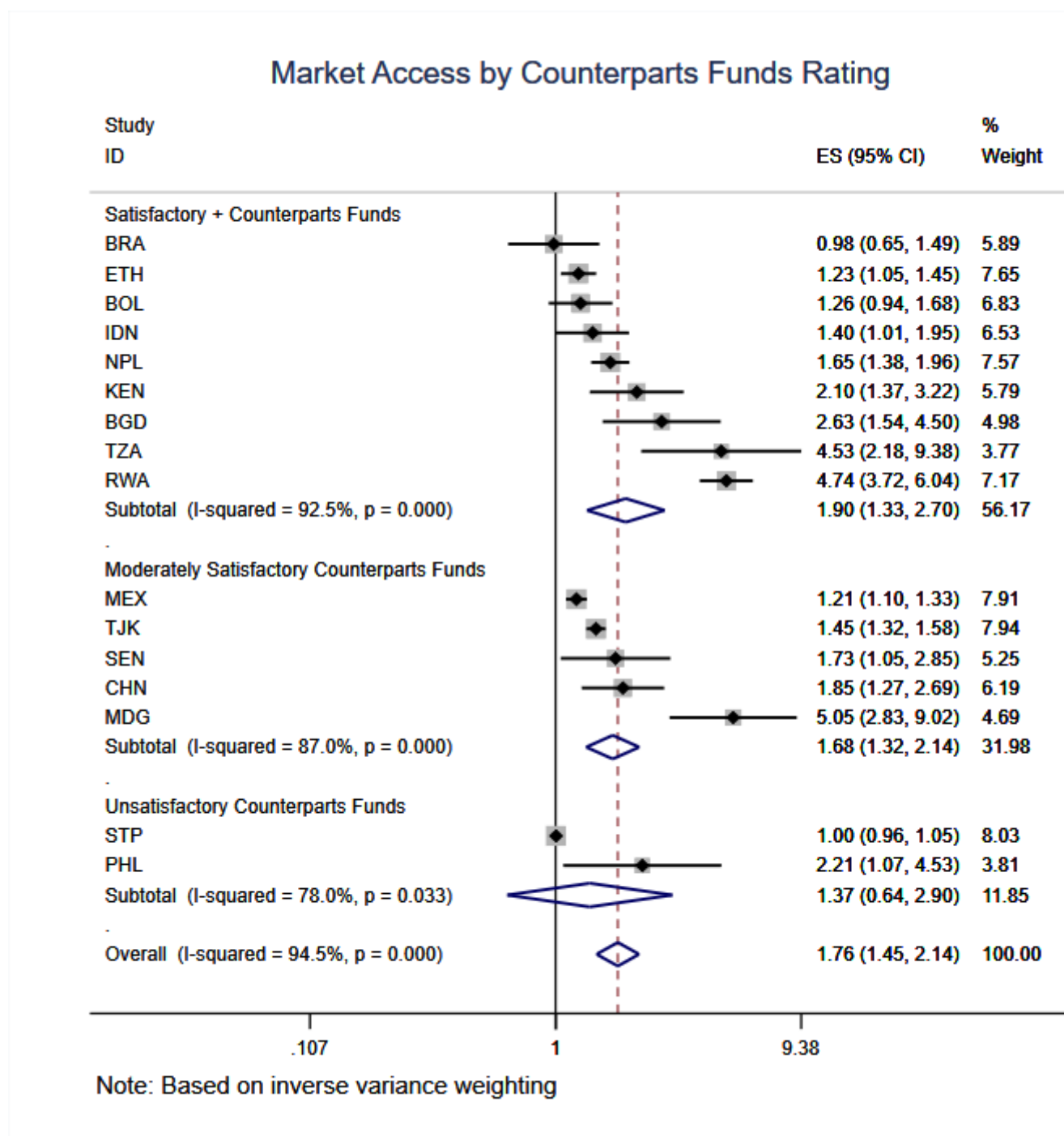
**Table 24**



**Table 25**

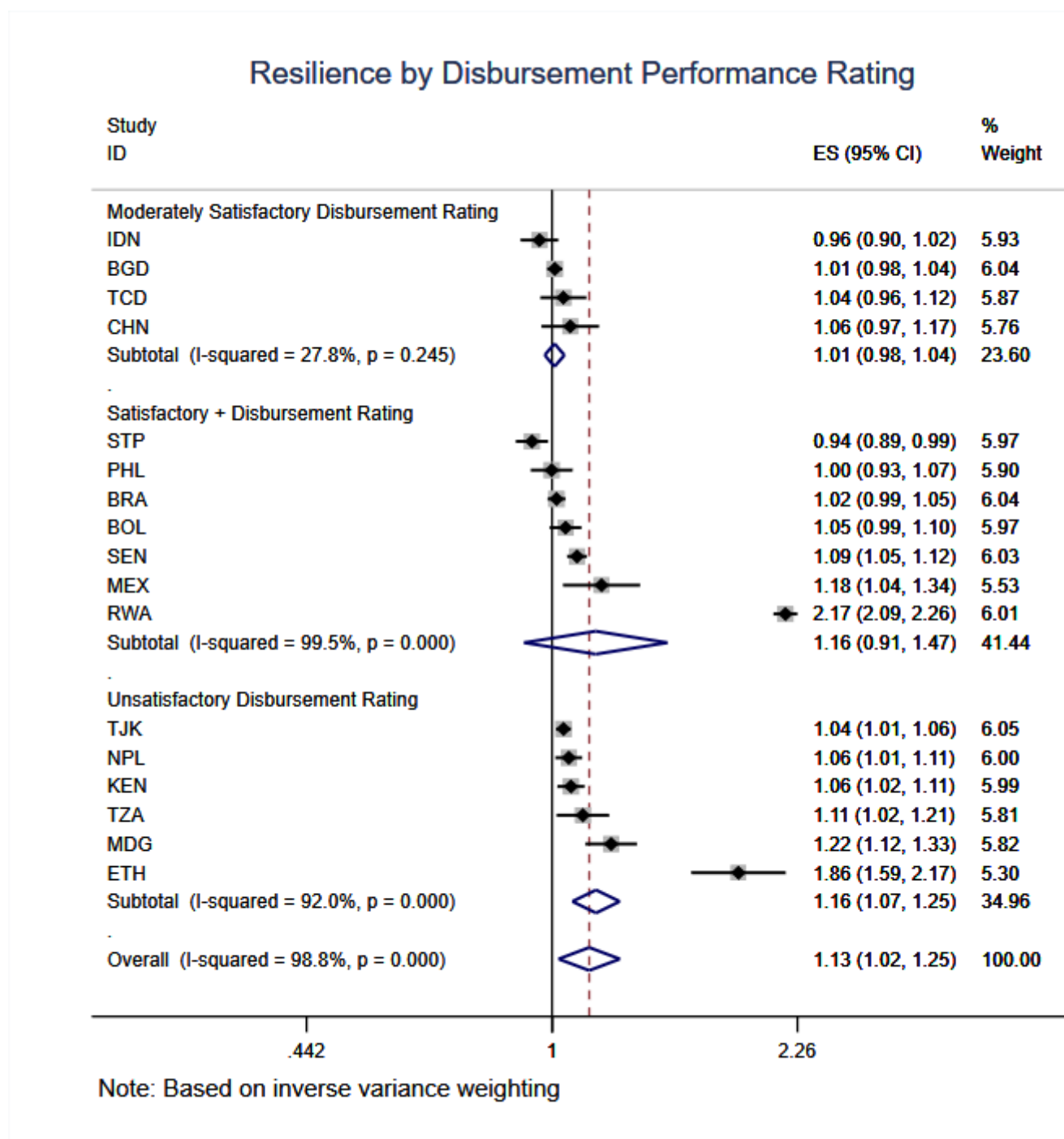


**Table 26**

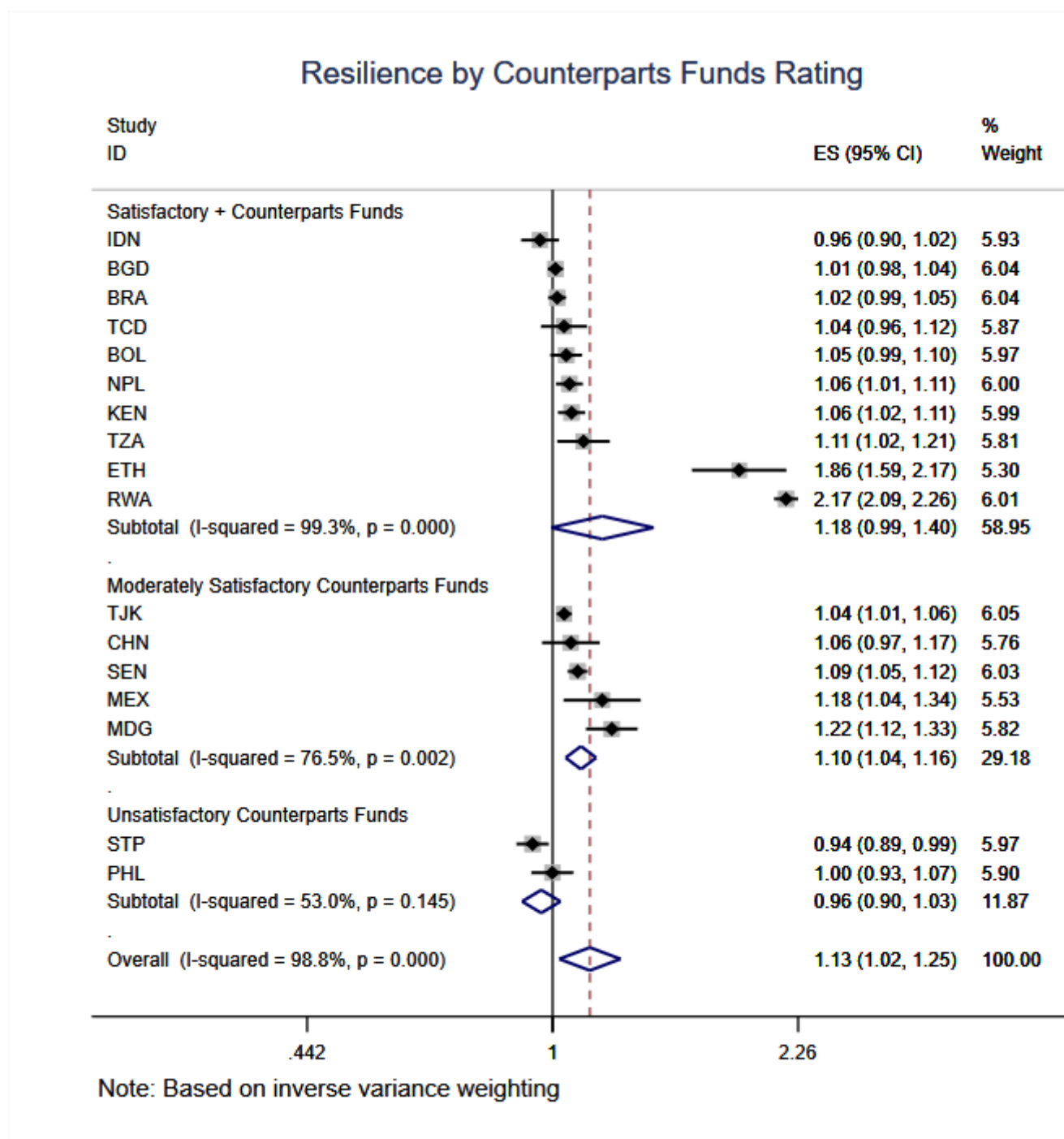


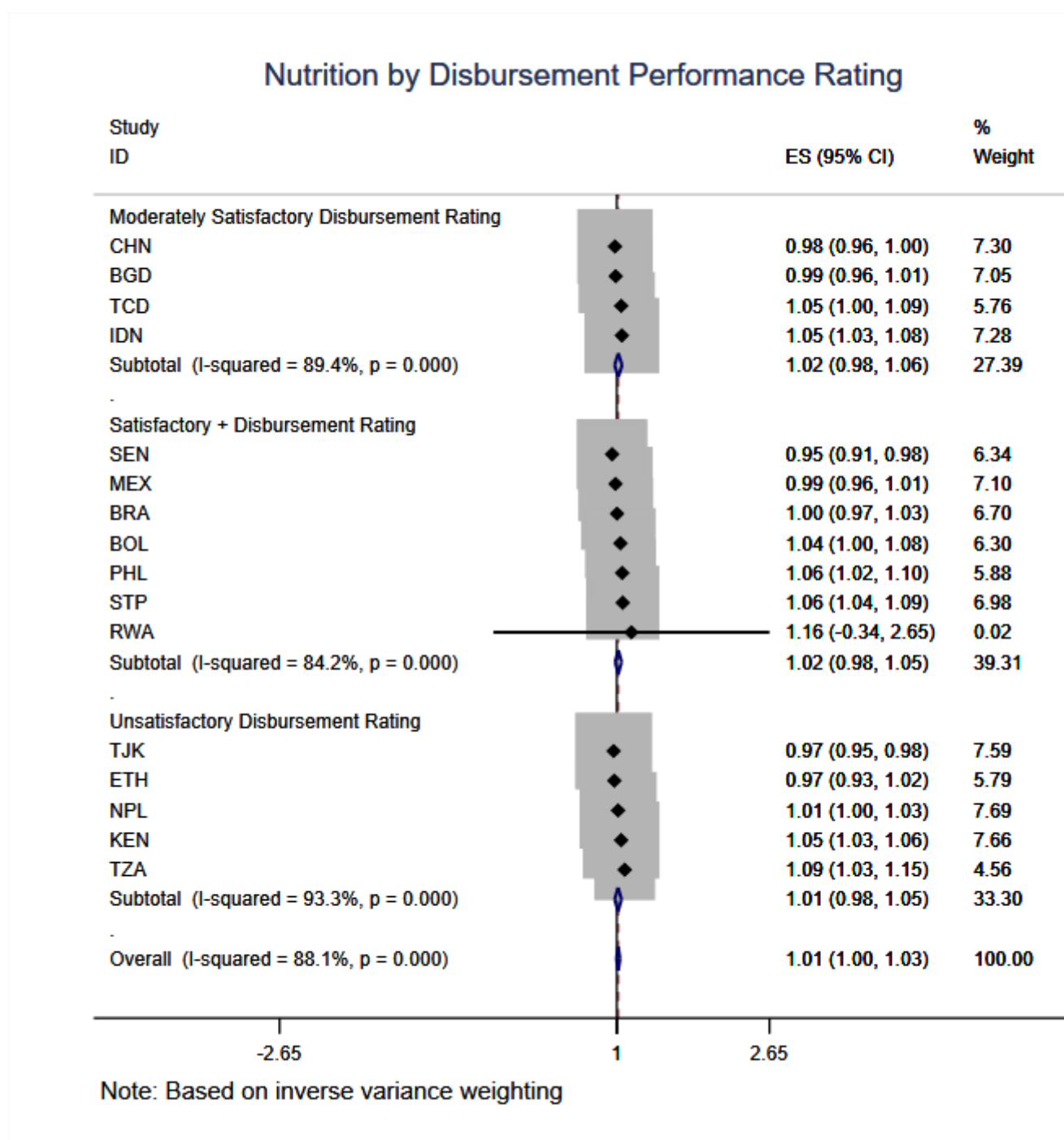


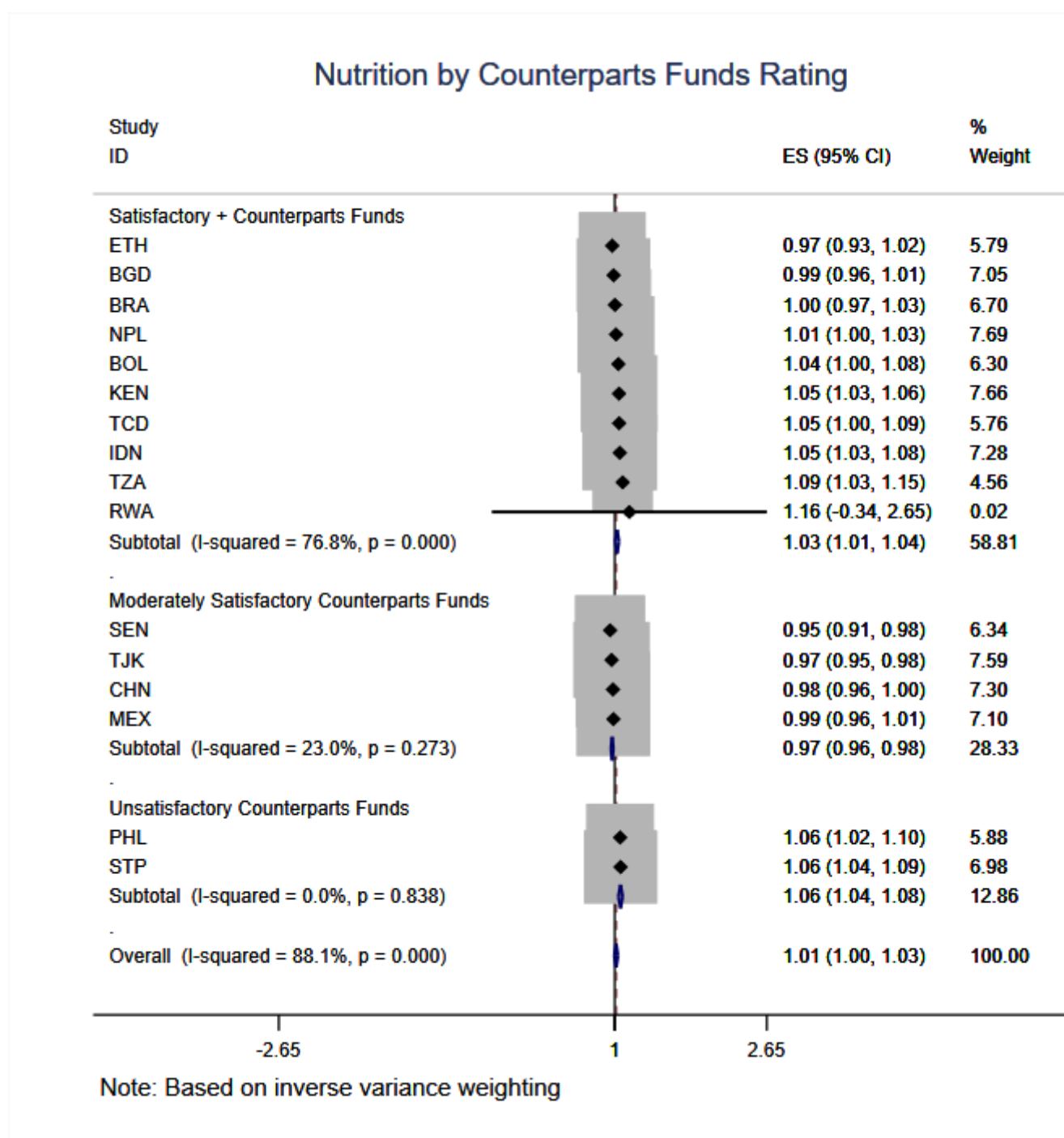
**Table 27**



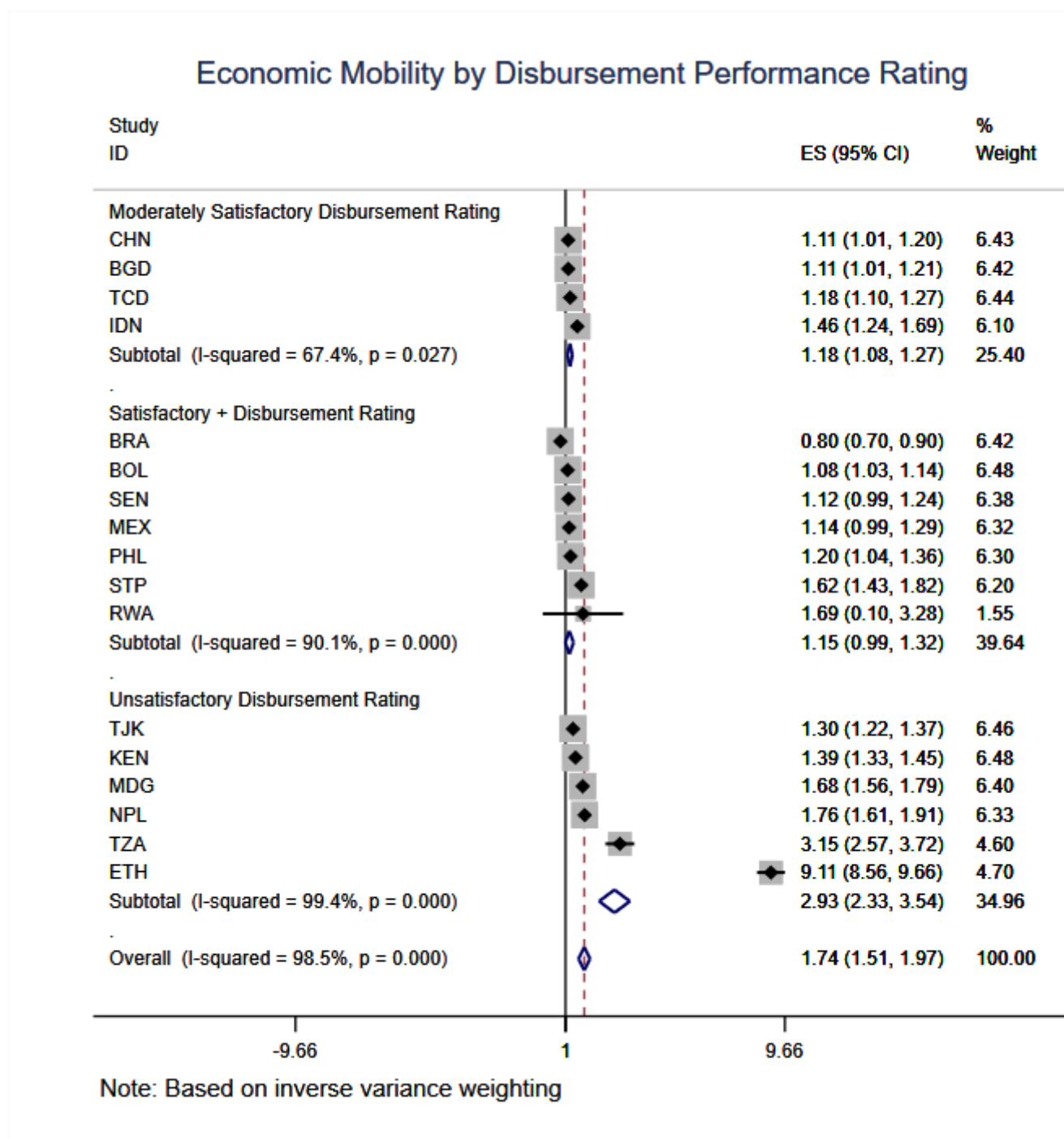
**Table 28**



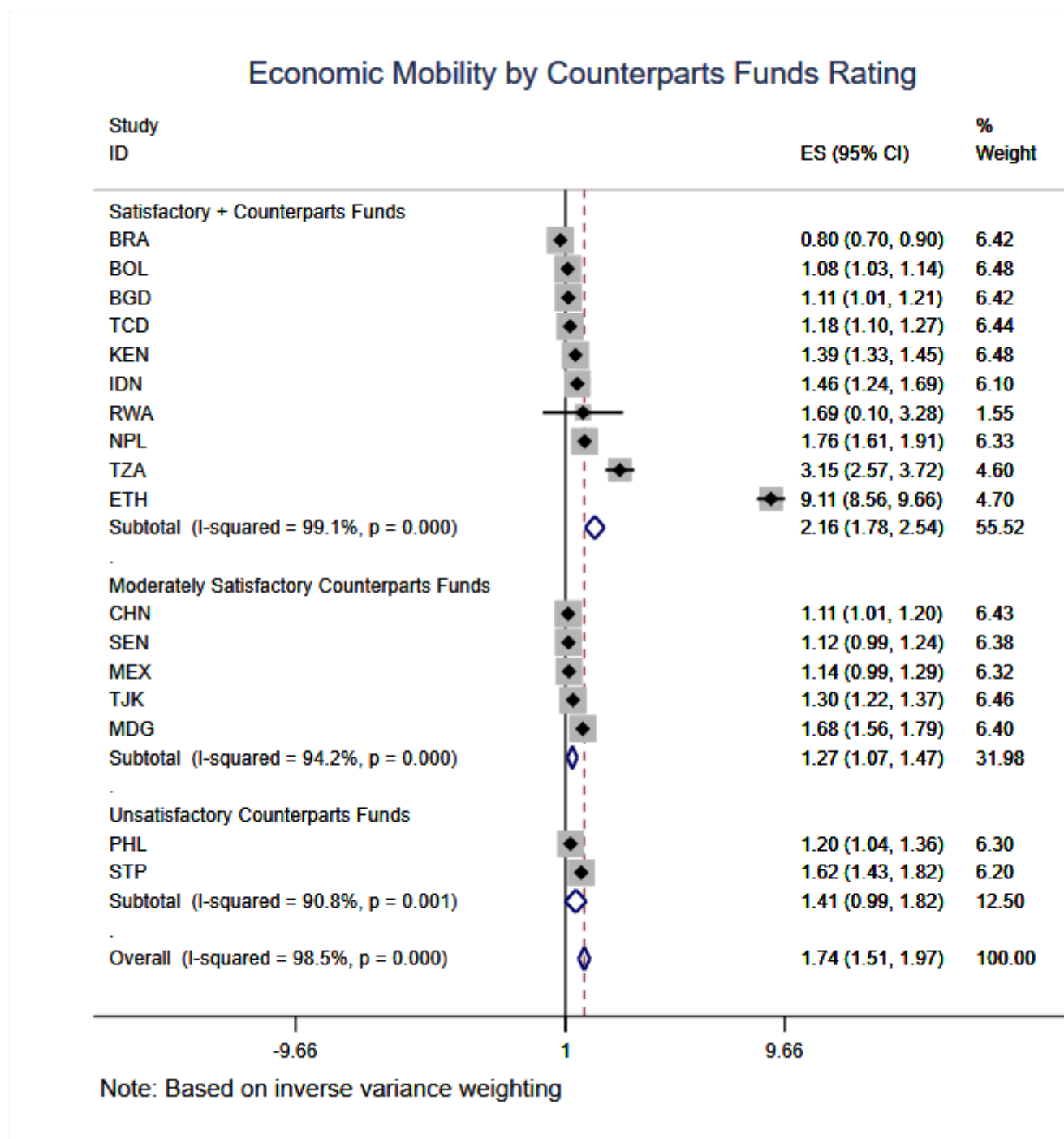
**Table 29**

**Table 30**

**Table 31**



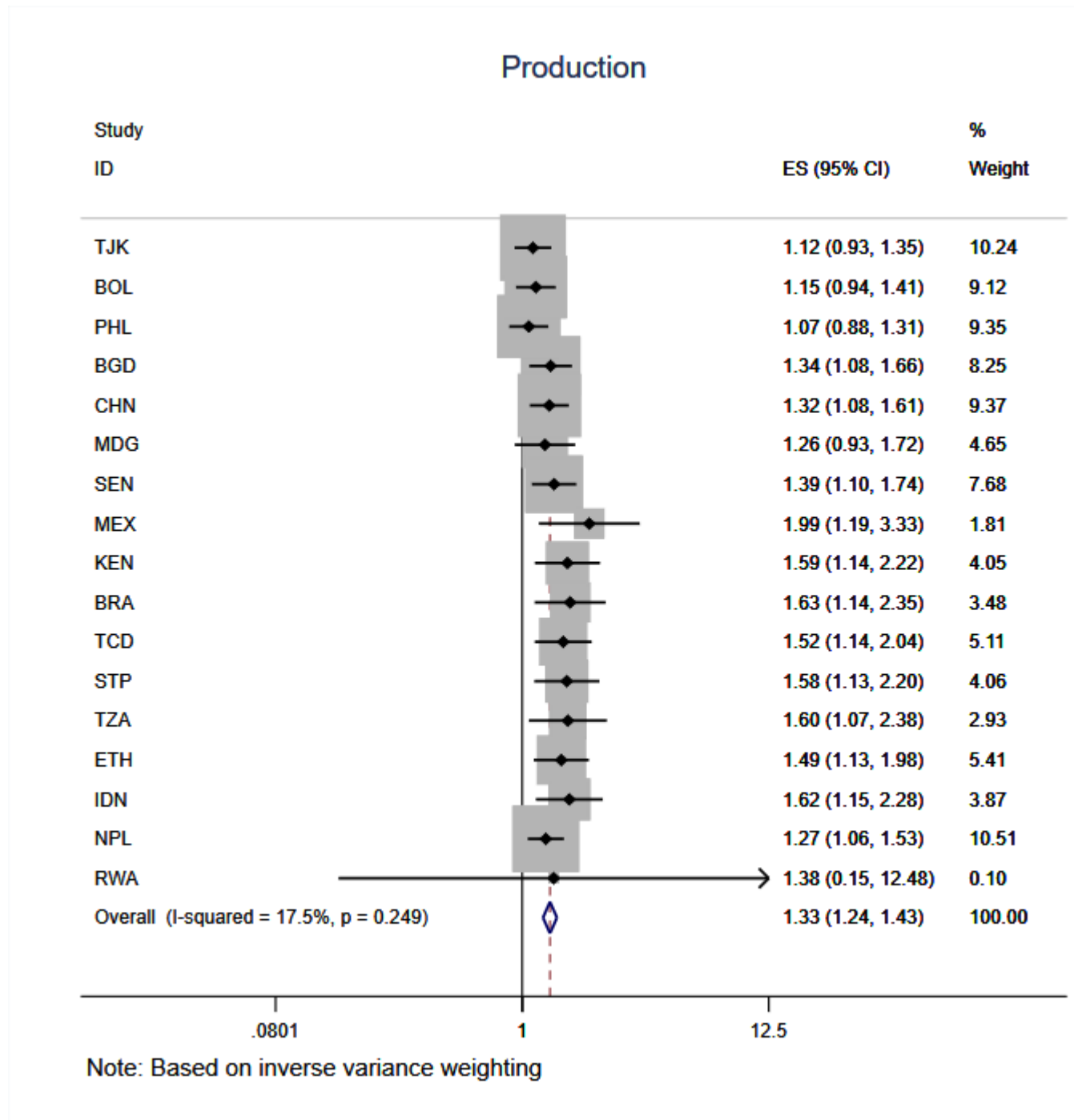
**Table 32**



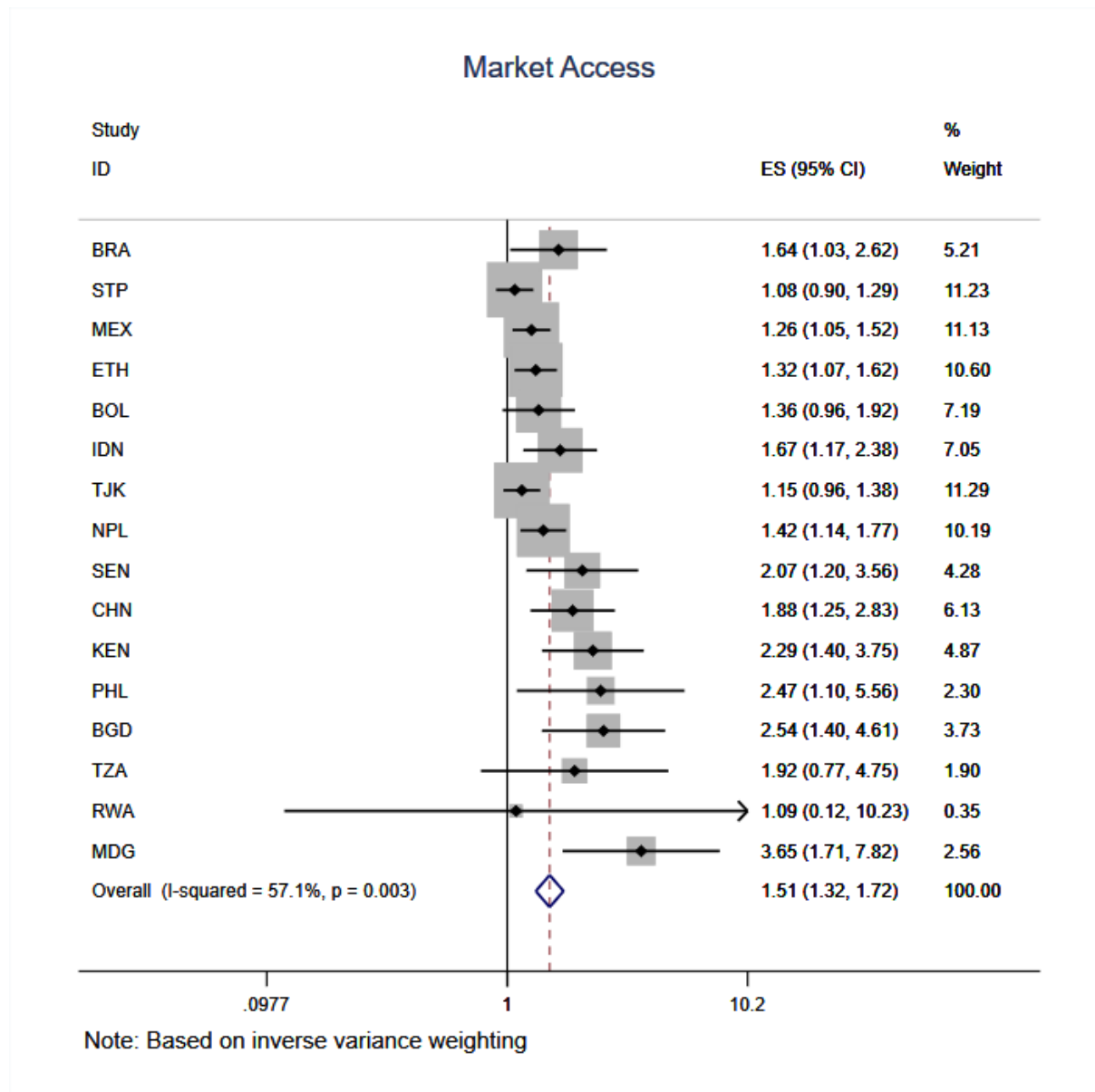
## Appendix - Annex III

### Sample correction a la Heckman

Table 33

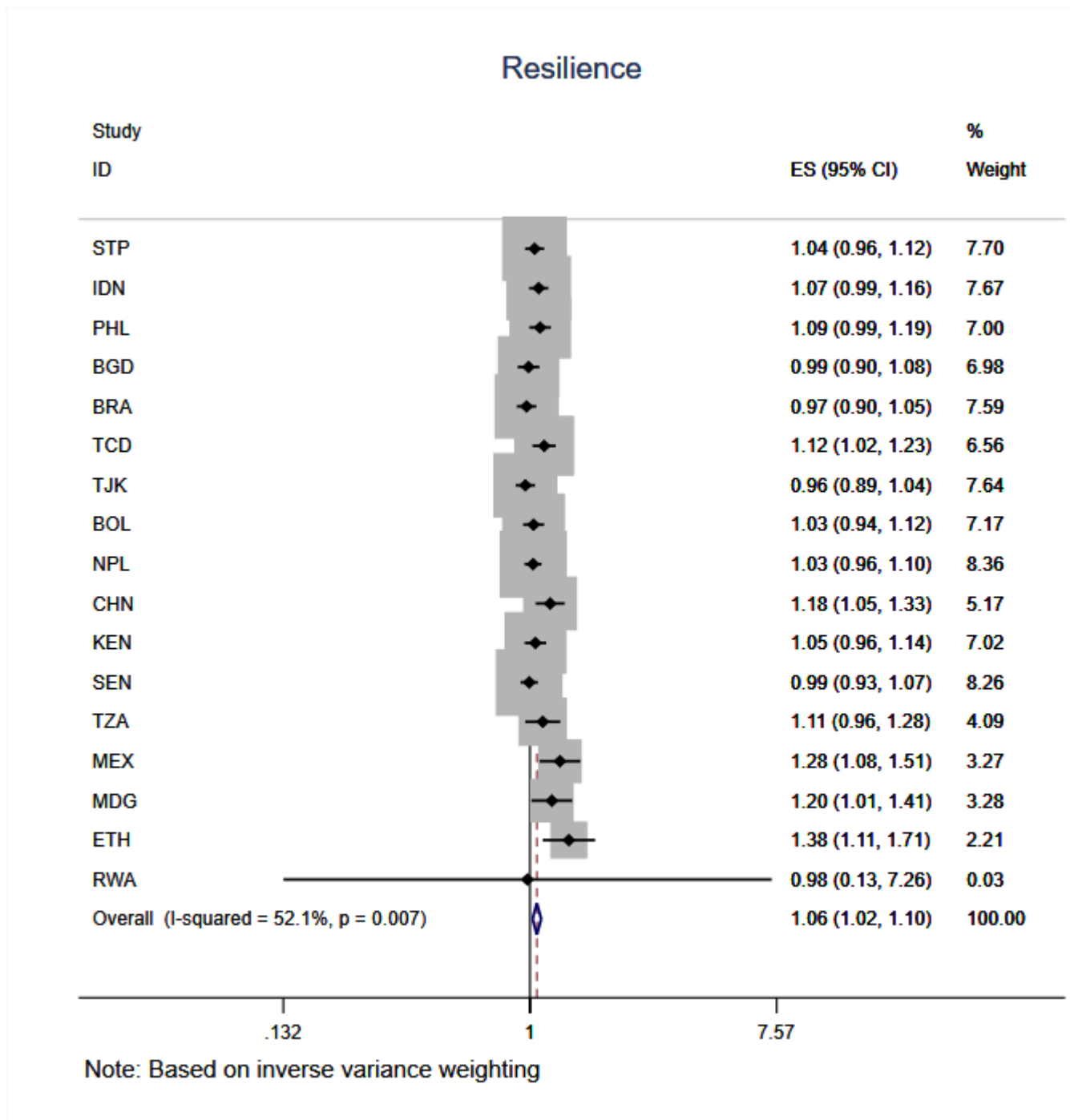


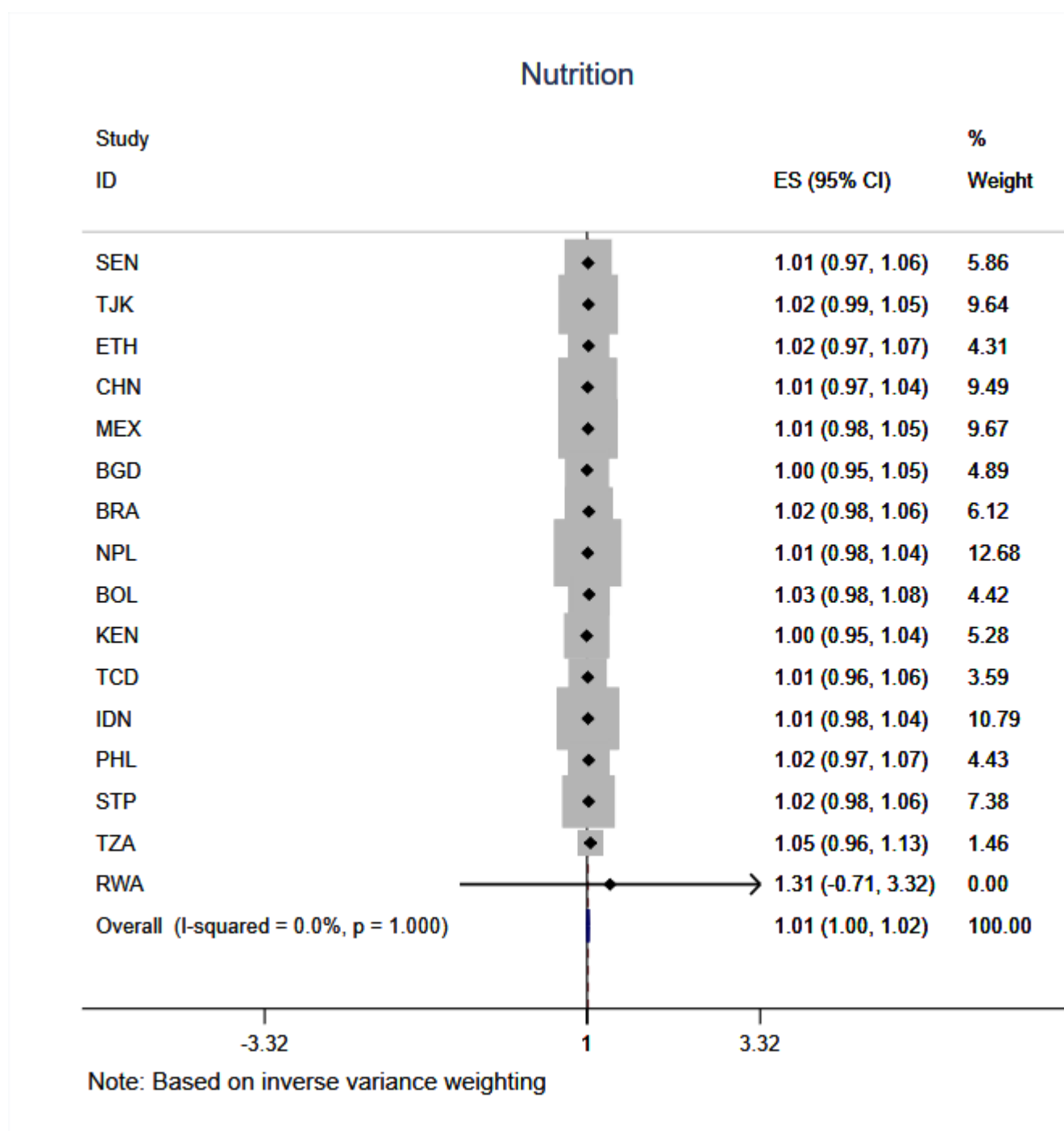
**Table 34**

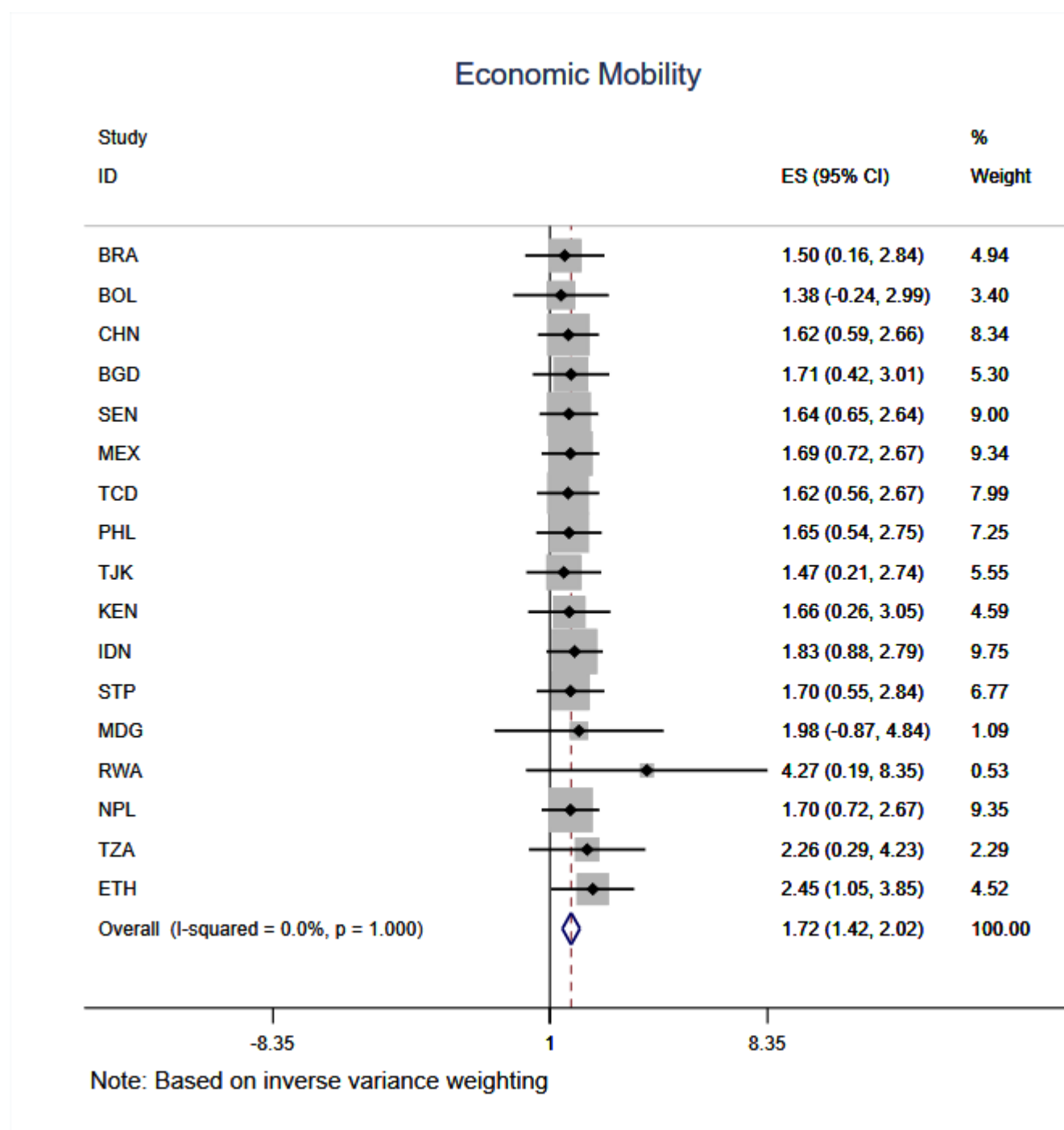




**Table 35**



**Table 36**

**Table 37**

## Appendix - Annex IV

### Sensitivity Analyses results with the trim-and-fill method

The trim and fill – adjusts for bias non-parametrically. Specifically, in order to investigate for the presence of small study effects and publication bias, visual representations such as funnel or contour enhanced funnel plots are employed.

A funnel plot shows effect sizes against measures of study precision e.g. standard error. The funnel plot is employed to explore visually publication bias or more precisely small study effect. The asymmetry is evidence and maybe the result of publication bias or may be because of other reasons (heterogeneity between studies).

The contour enhanced funnel plot, can help determine whether the asymmetry of the funnel plot is due to selection bias (e.g. publication bias). The contour lines correspond to certain levels of statistical significance. Publication bias is suspect when smaller studies are absent from the non-significant regions.

Tests for funnel-plot asymmetry are useful for detecting publication bias but are not able to estimate the impact of this bias on the final meta-analysis results. The nonparametric trim-and-fill method of Duval and Tweedie (2000a, 2000b) provides a way to assess the impact of missing studies because of publication bias on the meta-analysis. It evaluates the amount of potential bias present in meta-analysis and its impact on the final conclusion. This method is typically used as a sensitivity analysis to the presence of publication bias.

Results from the Trim-and-fill method are presented in Table 38 which summarizes the original results for the meta-analysis (the observed effect size – ES) along with the imputed one from the trim and fill results (observed plus imputed ES). The full set of tables are **Error! Reference source not found.** to Table 48.

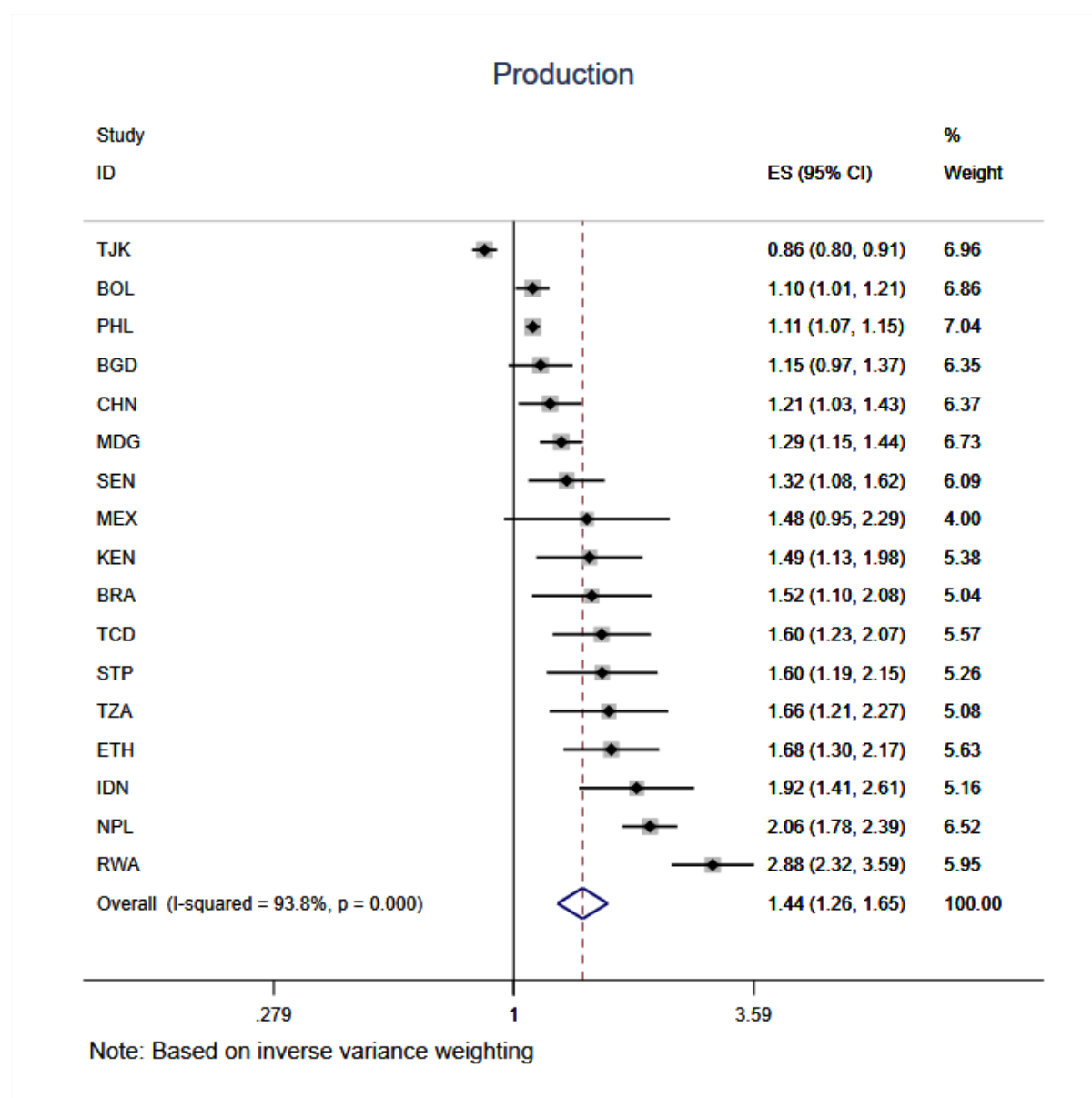
The table shows that adjusted results remain largely positive and sometimes unaltered for some domains. Bias might affect only three coefficients: Economic mobility (1.74 vs. 1.38 equivalent to 74% and 38% respectively), Market access (1.76 vs. 1.38 equivalent to 76% and 38% respectively); and Resilience indicators (1.13 vs. 1.03 equivalent to 13% and 3%, respectively). However a known limitation of Trim-and-Fill is that it can correct for publication bias that does not exist, underestimating effect sizes (Terrin et al 2003). Recommendations from recently published literature, (Simonsohn et al 2014) argued against the use of such method<sup>23</sup>.

---

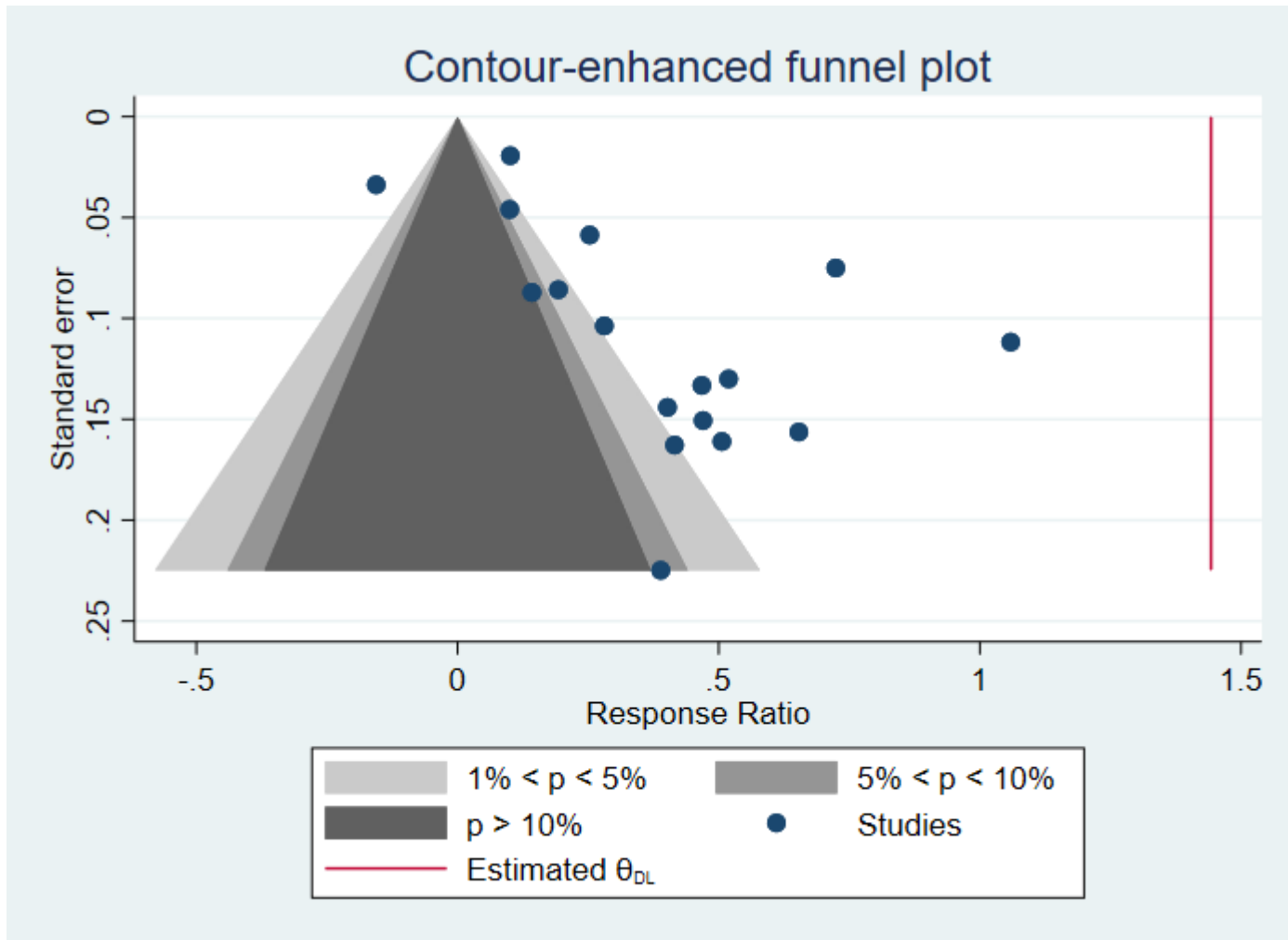
<sup>23</sup> Simonsohn, Uri and Nelson, Leif D. and Simmons, Joseph P., P-Curve and Effect Size: Correcting for Publication Bias Using Only Significant Results (April 27, 2014). Available at SSRN: <https://ssrn.com/abstract=2377290> or <http://dx.doi.org/10.2139/ssrn.2377290>

**Table 38: Results from the Trim and Fill method**

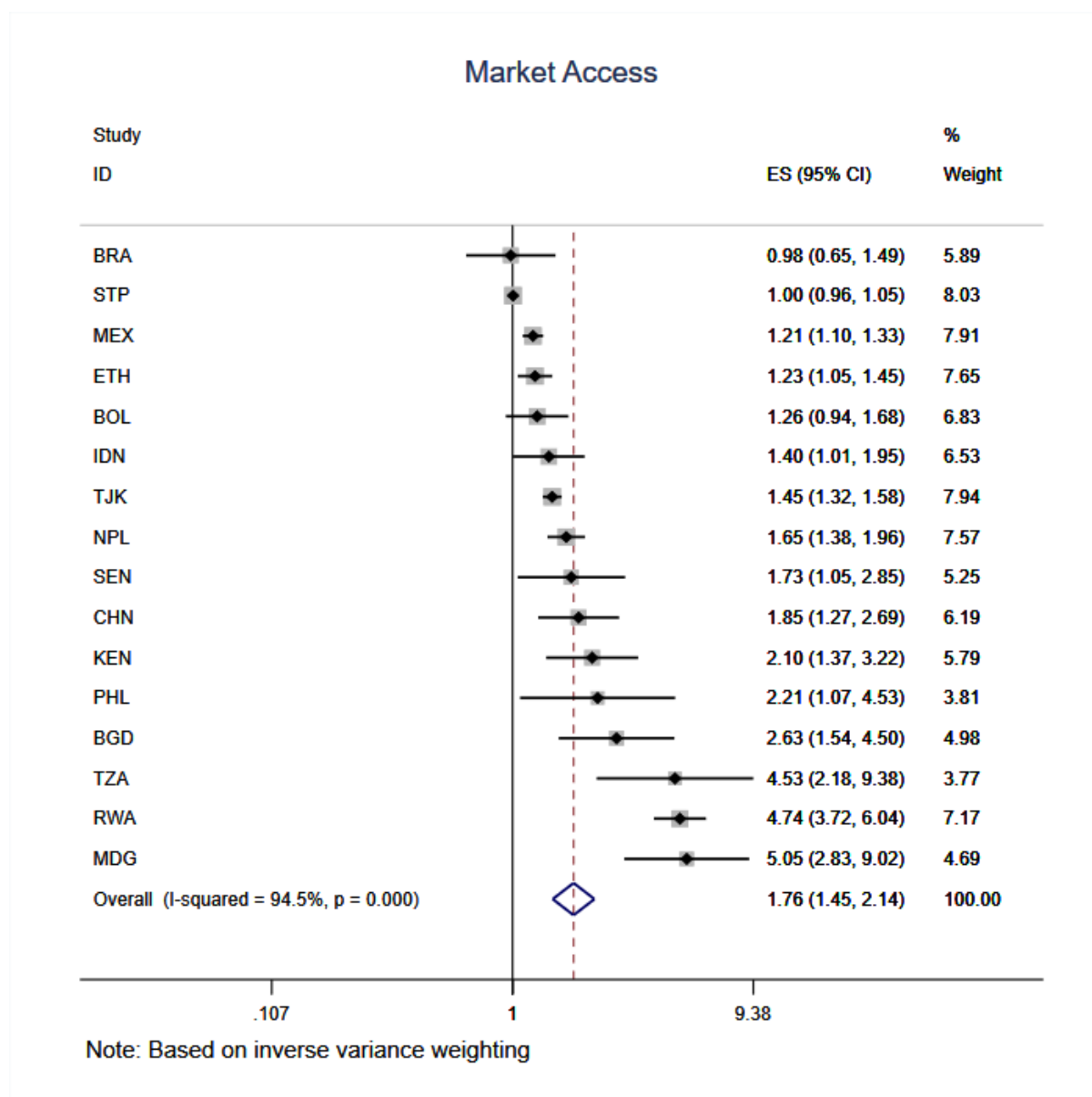
<b>Production</b>				
	N. projects	ES	Lower CI	Upper CI
<b>Observed</b>	17	1.44	1.26	1.65
<b>Observed+Imputed</b>	17	1.44	1.26	1.65
<b>Market Access</b>				
	N. projects	ES	Lower CI	Upper CI
<b>Observed</b>	16	1.76	1.45	2.14
<b>Observed+Imputed</b>	21	1.38	1.13	1.67
<b>Resilience</b>				
	N. projects	ES	Lower CI	Upper CI
<b>Observed</b>	17	1.13	1.02	1.25
<b>Observed+Imputed</b>	20	1.04	0.91	1.18
<b>Nutrition</b>				
	N. projects	ES	Lower CI	Upper CI
<b>Observed</b>	16	1.01	1	1.03
<b>Observed+Imputed</b>	17	1.01	1	1.03
<b>Economic Mobility</b>				
	N. projects	ES	Lower CI	Upper CI
<b>Observed</b>	17	1.74	1.51	1.97
<b>Observed+Imputed</b>	18	1.38	1.1	1.67

**Table 39**

**Table 40**

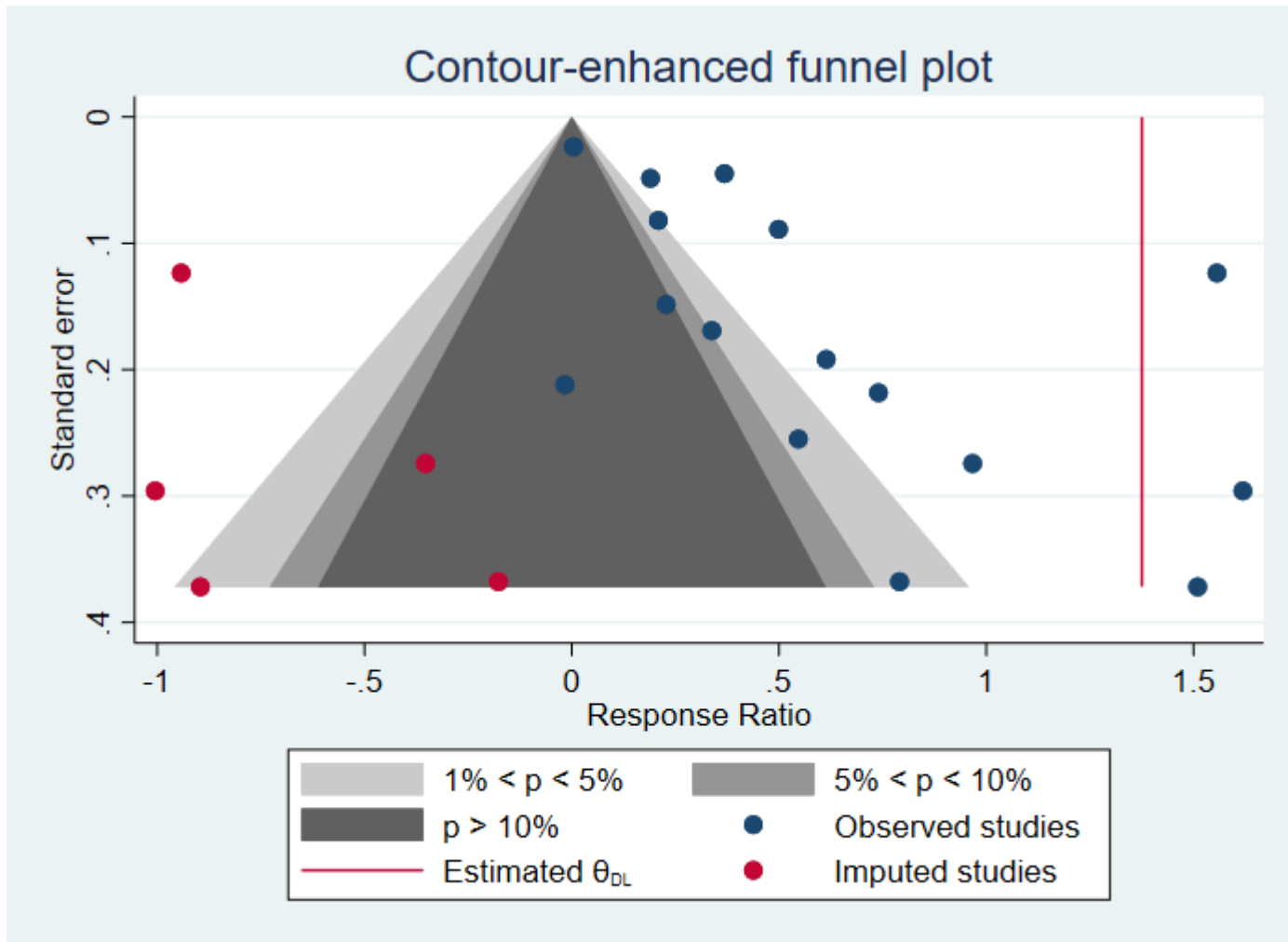


**Table 41**

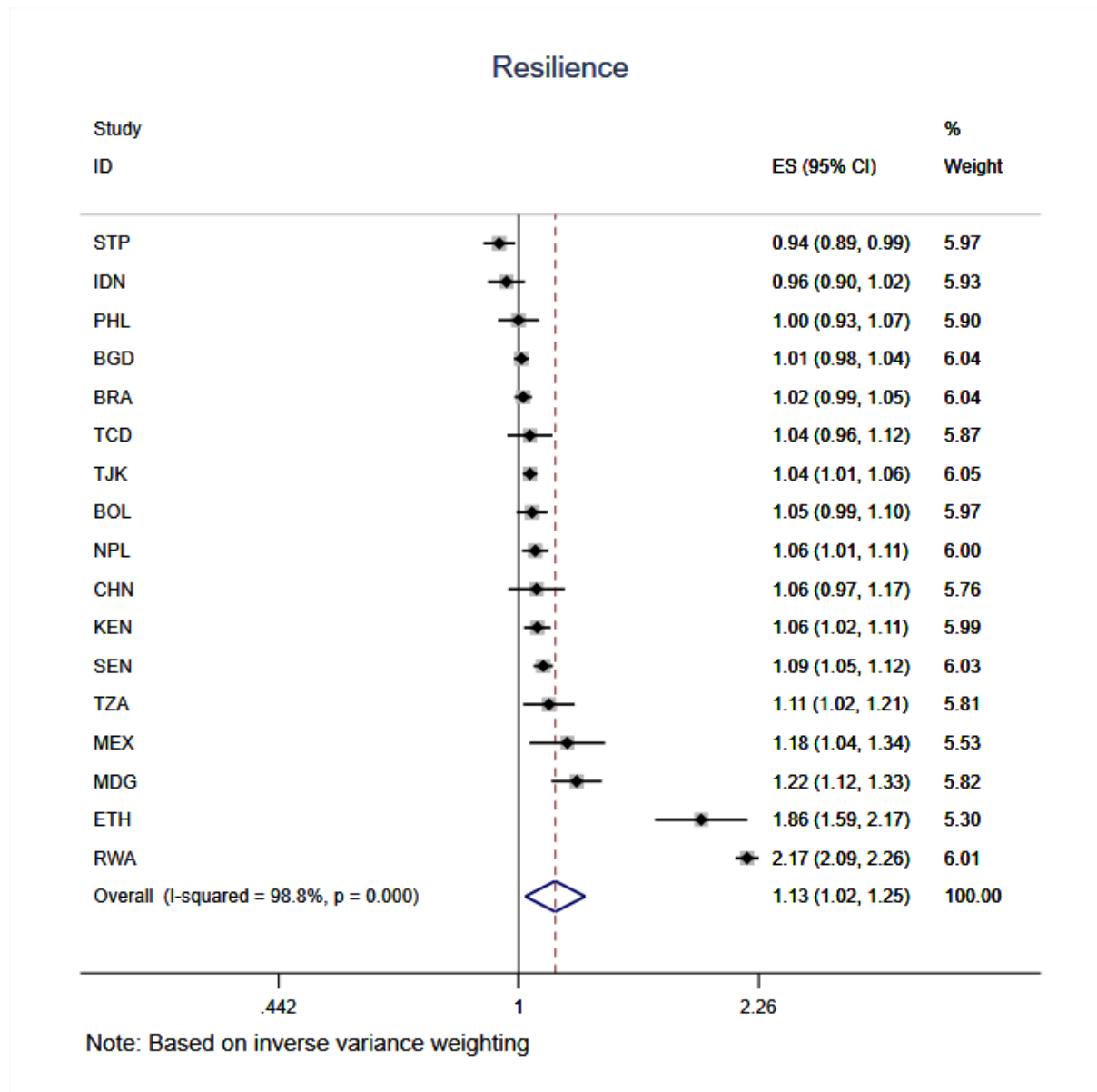




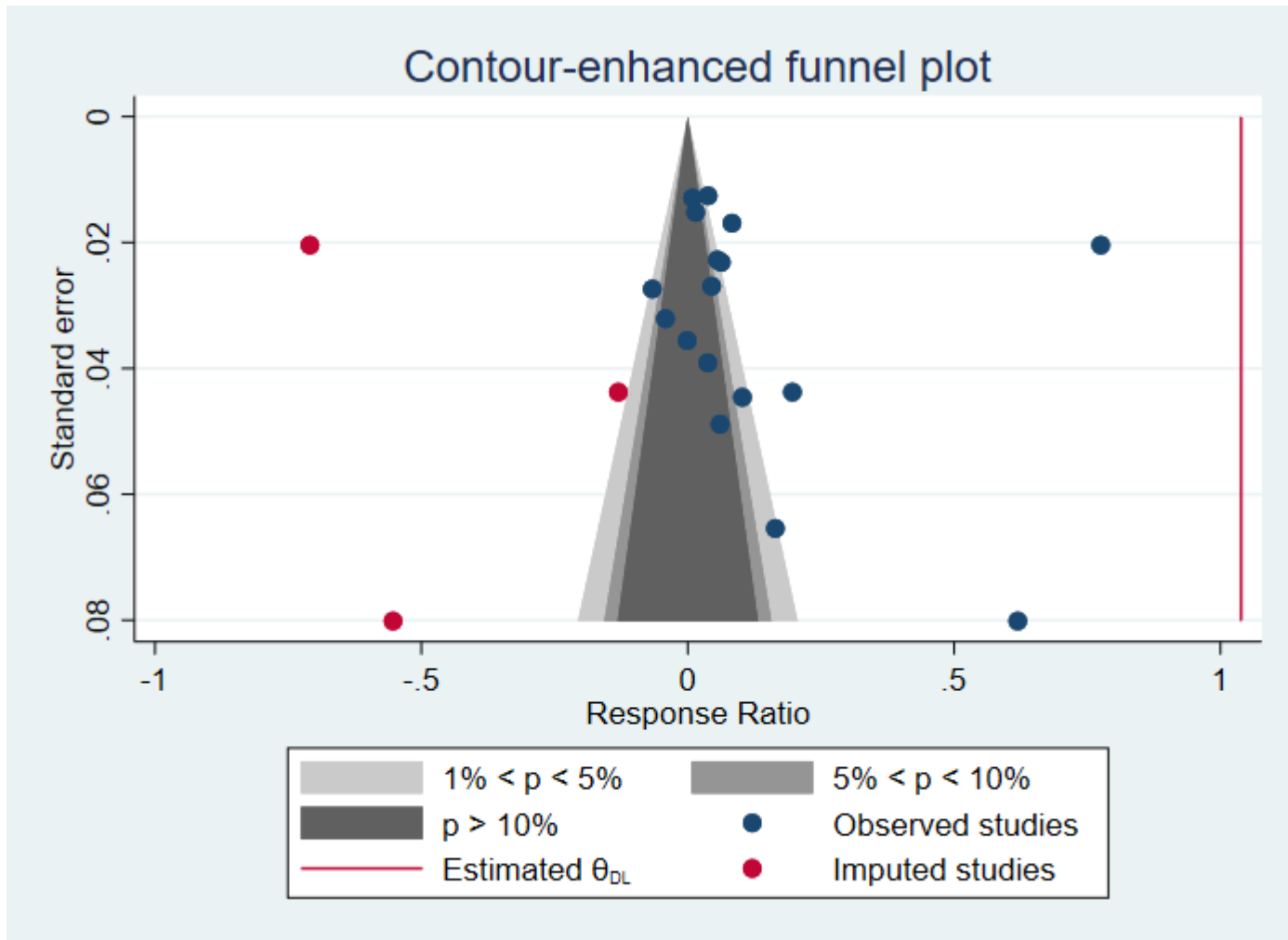
**Table 42**

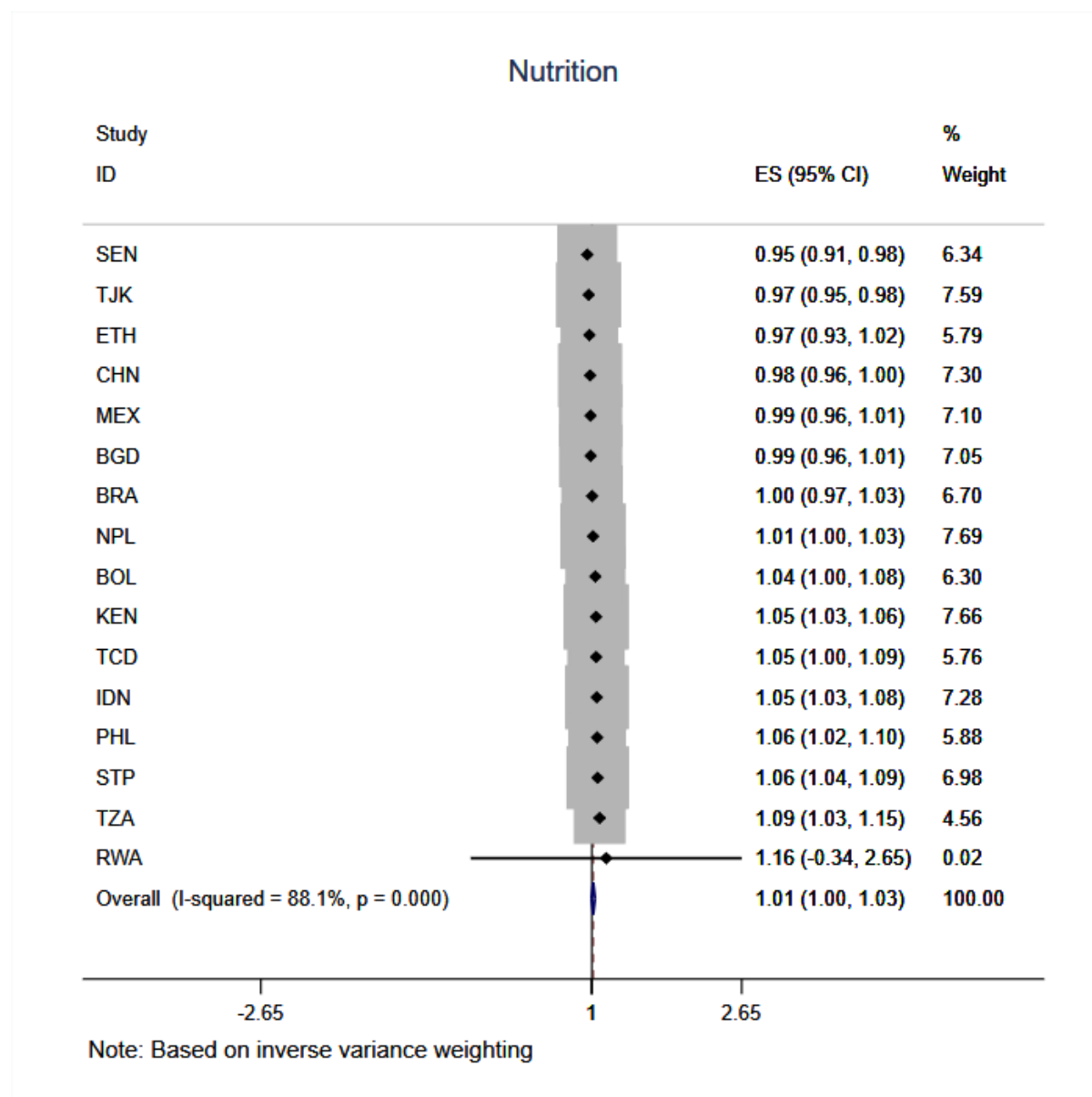


**Table 43**

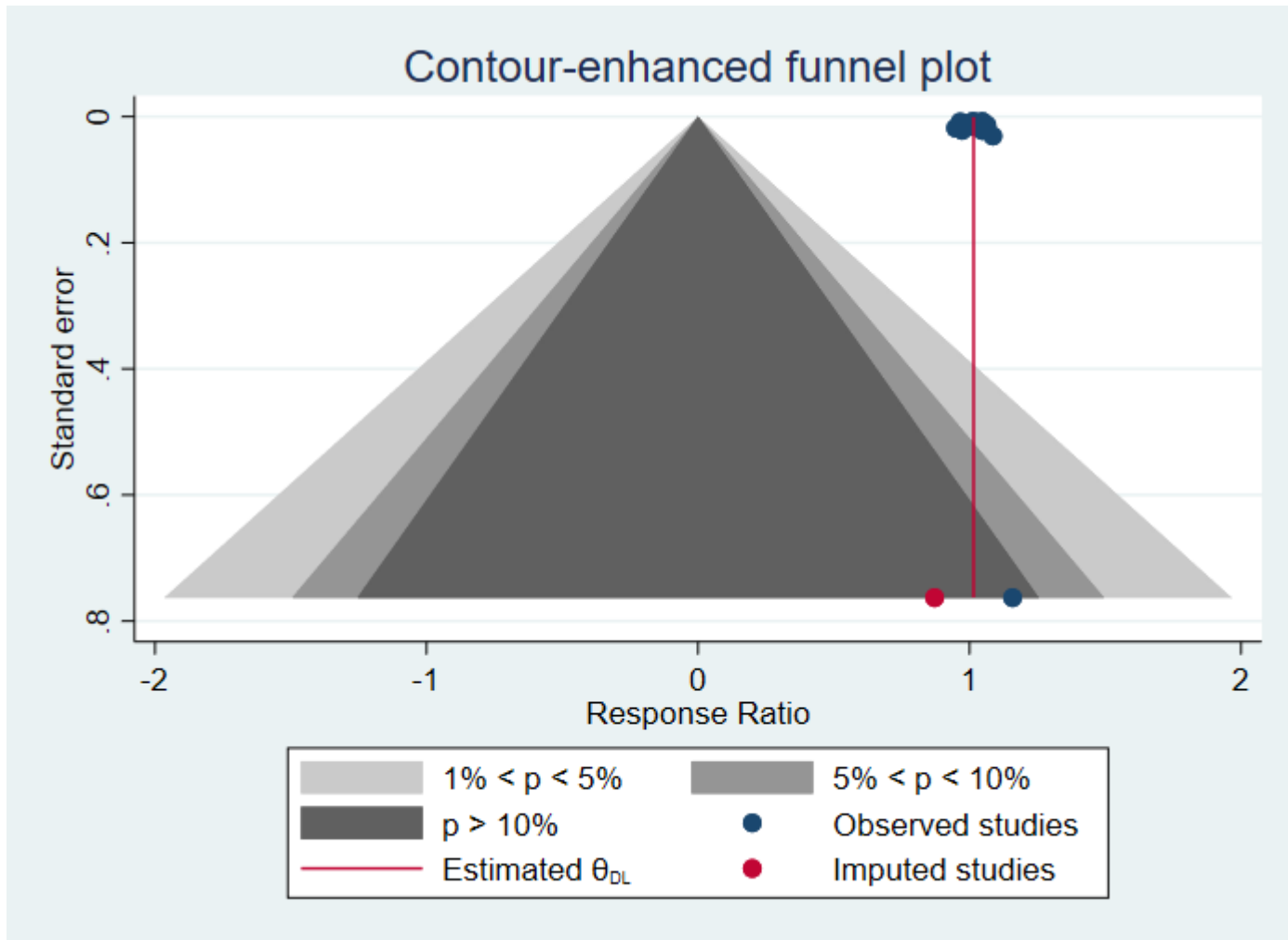


**Table 44**

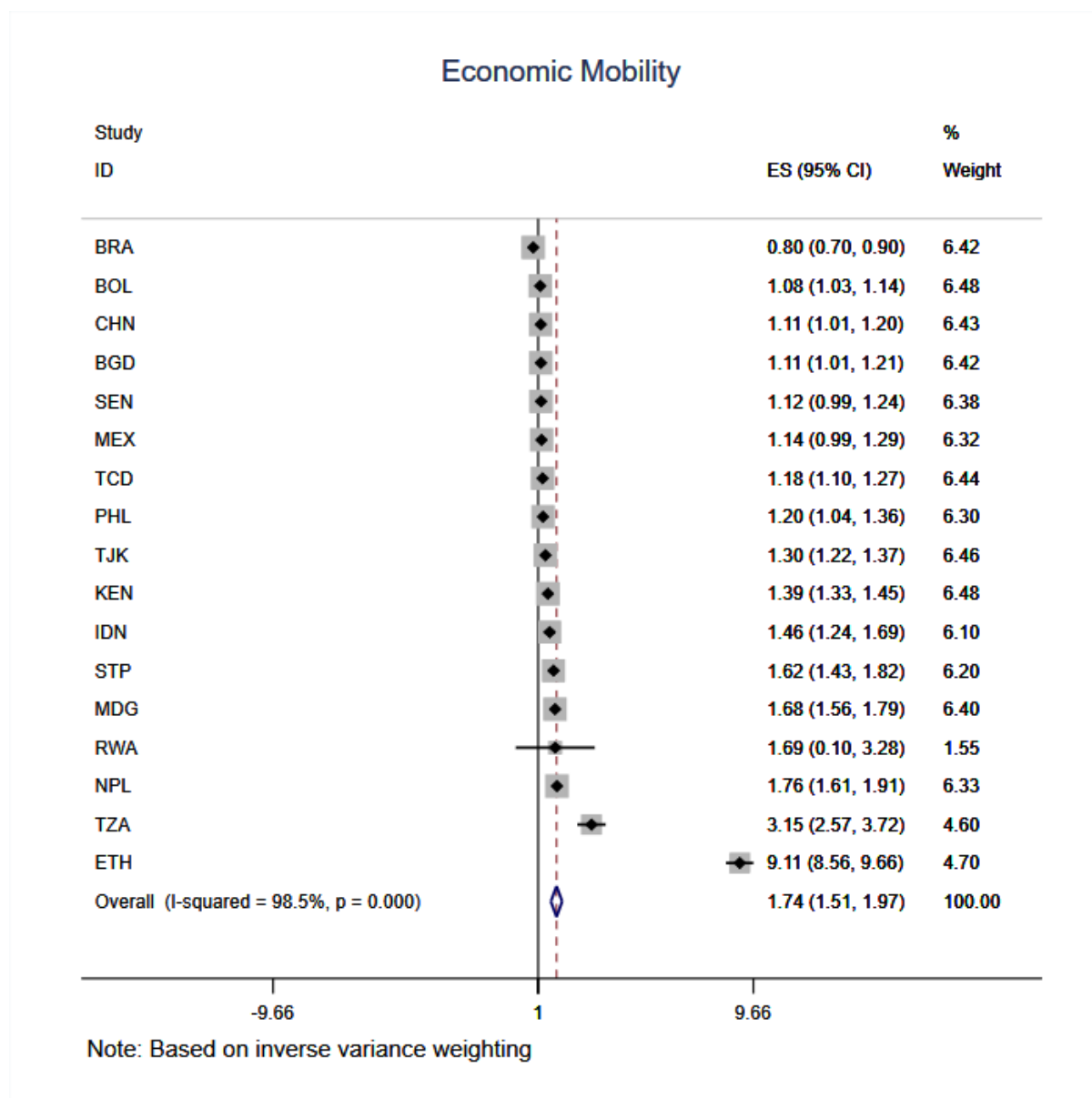


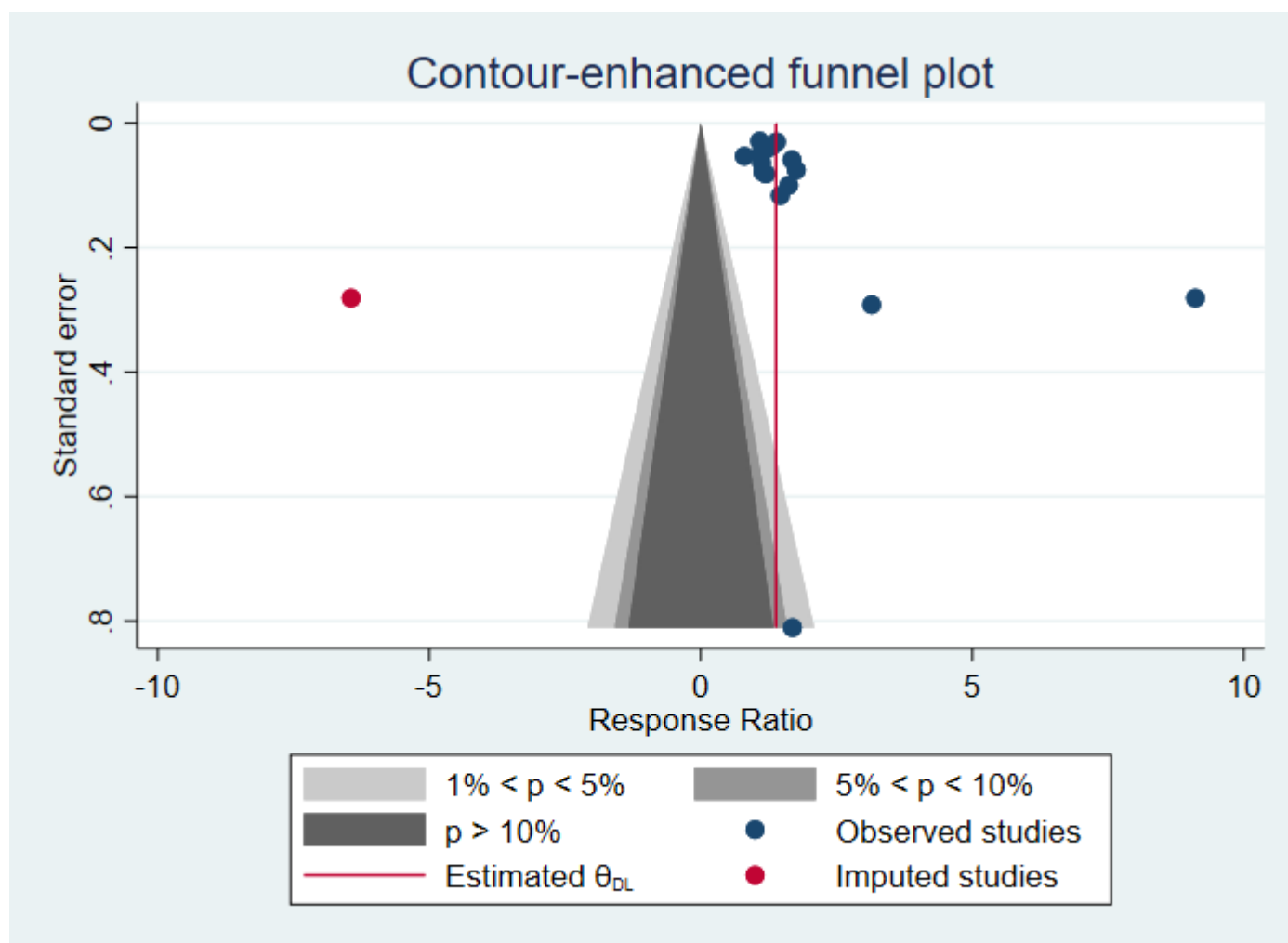
**Table 45**

**Table 46**



**Table 47**



**Table 48**

## References

- Andrews, Isaiah and Emily Oster. 2018. "Weighting for External Validity." National Bureau of Economic Research Working Paper 23826. <http://www.nber.org/papers/w23826>.
- Banerjee, Amitav and Chaudhury, Suprakash. 2010. "Statistics without tears: Populations and samples." *Industrial Psychiatry Journal*, 19(1): 60–65.
- Bown, M.J. and A.J. Sutton. 2010. "Quality Control in Systematic Reviews and Meta-analyses." *European Journal of Cardiovascular and Endovascular Surgery*, 40: 669-677.
- Breslow NE, Day NE. 1980. *Statistical Methods in Cancer Research: Vol. I - The Analysis of Case-Control Studies*. Lyon: International Agency for Research on Cancer.
- Copas, John and Jian Qing Shi. 2000. "Meta-analysis, funnel plots, and sensitivity analysis." *Biostatistics*, 1(3): 247-262.
- Clarke, Kevin A, Randall W. Stone. 2019. "The Unobserved IMF." *The Political Economy of International Organizations*. [https://www.peio.me/wp-content/uploads/2019/01/PEIO12\\_paper\\_109.pdf](https://www.peio.me/wp-content/uploads/2019/01/PEIO12_paper_109.pdf)
- Cortes, Corinna; Mohri, Mehryar; Riley, Michael; Rostamizadeh, Afshin (2008). *Sample Selection Bias Correction Theory (PDF)*. *Algorithmic Learning Theory. Lecture Notes in Computer Science*. 5254. pp. 38–53.

- Cortes, Corinna; Mohri, Mehryar (2014). "Domain adaptation and sample bias correction theory and algorithm for regression" (PDF). *Theoretical Computer Science*. 519: 103–126.
- Delgado-Rodríguez M, Llorca J. "Bias" *Journal of Epidemiology & Community Health* 2004;58:635-641.
- Greenland, S. (2005) Multiple-bias modelling for analysis of observational data. *J. R. Statist. Soc.A*,168, 267–291
- Heckman, James J., Sergio Urzua and Edward Vytlacil. 2006. "Understanding Instrumental Variables In Models With Essential Heterogeneity," *Review of Economics and Statistics*, 88(3, Aug): 389-432
- Henmi M, Copas JB. 2010. Confidence intervals for random effects meta-analysis and robustness to publication bias. *Statistics in Medicine* 29: 2969-2983. doi: 10.1002/sim.4029
- Lin L (2018) Bias caused by sampling error in meta-analysis with small sample sizes. *PLoS ONE* 13(9): e0204056. <https://doi.org/10.1371/journal.pone.0204056>
- Mathur M.B & VanderWeele Tyler J. (2019): Sensitivity Analysis for Unmeasured Confounding in Meta-Analyses, *Journal of the American Statistical Association*, DOI: 10.1080/01621459.2018.1529598
- Mavridis D, Salanti G. How to assess publication bias: funnel plot, trim-and-fill method and selection models. *Evidence-Based Mental Health* 2014;17:30.
- Oster, Emily. 2019. "Unobservable Selection and Coefficient Stability: Theory and Evidence." *Journal of Business & Economic Statistics*, 37(2):187-204.
- Shi, Linyu and Lifeng Lin. 2019. "The trim-and-fill method for publication bias: practical guidelines and recommendations based on a large database of meta-analyses." *Medicine*, 98(23): 1-11.
- Stuart, Elizabeth A., Stephen R. Cole, Catherine P. Bradshaw, and Philip J. Leaf. 2011. "The use of propensity scores to assess the generalizability of results from randomized trials." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 174(2): 369-386.
-