**BAB I**

**PENDAHULUAN**

1. Latar Belakang Masalah

Salah satu penilaian yang dilaksanakan pemerintah dalam bentuk tes formatif adalah ujian nasional dalam PP No 19 Tahun 2005 tentang Standar Nasional Pendidikan, BAB 10 pasal 63 disebutkan standar penilaian pendidikan yang diklasifikasikan kepada 3 kriteria yaitu: (1) penilaian hasil belajar oleh pendidik, (2) penialaian hasil belajar oleh satuan pendidikan, dan (3) penilaian hasil belajar oleh pemerintah. Khusus pada kriteria ketiga, dalam PP ini dijelaskan secara lebih spesifik lagi dalam pasal 66 butir 1 dinyatakan bahwa penilaian hasil belajar sebagaimana dimaksud dalam pasal 63 ayat 1 butir c bertujuan untuk menilai pencapaian kompetensi lulusan secara nasional pada mata pelajaran tertentu dalam kelompok mata pelajaran ilmu pengetahuan teknologi yang dilakukan dalam bentuk Ujian Nasional (UN).

Berpedoman pada penjelasan di atas, maka dapat disimpulkan bahwa ujian nasional (UN) merupakan salah satu instrumen dalam sistem pendidikan di Indonesia guna mengukur kompetensi siswa, keberhasilan guru mengajar dan dijadikan standar mutu kelulusan (*quality control*). Harapan lain dari pelaksanaan UN adalah agar pelajar dan pembelajar serta pengelola pendidikan dapat meningkatkan kinerja dan kualitasnya. Sebagai alat ukur tentunya dikatakan baik apabila tes tersebut mampu memberikan hasil yang cermat, akurat dan berkualitas.

Ujian nasional yang dilaksanakan terdiri atas beberapa paket soal yang menjadi masalah apakah paket-paket soal tersebut dapat mengukur secara obyektif terhadap peserta didik. Agar peserta didik yang satu dengan peserta didik yang lainnya tidak ada yang dirugikan atau diuntungkan dari paket soal yang berbeda maka dalam evaluasi pembelajaran aspek keadilan merupakan salah satu prinsip yang penting, artinya peserta didik yang menghadapi ujian nasional di berbagai wilayah Indonesia dan waktu yang berbeda harusnya mendapatkan perlakuan yang adil. Dengan demikian apakah soal ujian nasional mempunyai tingkat keseteraan yang sama antara paket yang satu dengan yang lainnya.

Salah satu pendapat bahwa “Dalam pelaksanaan evaluasi yang menggunakan beberapa perangkat tes yang berbeda dan mengukur hal yang sama perlu dilakukan penyetaraan dari beberapa perangkat tes tersebut, karena dengan adanya penyetaraan dapat dijamin keadilan bagi peserta tes” (Suryabrata, 1987: 132). Agar peserta didik mendapatkan penilaian yang adil dalam menghadapi soal tes yang berbeda paket. Miyatun & Mardapi (2000) menyatakan bahwa “Tingkat kesetaraan perangkat tes yang berbeda akan dapat diketahui melalui preses penyetaraan”. Selanjutnya Hambleton & Swaminathan (1985) dalam Salasi (1996) mengingatkan bahwa sekalipun perangkat tes yang disusun berdasarkan kisi-kisi yang sama, jarang sekali bahkan hampir tidak ditemukan perangkat tes yang benar-benar setara dalam sebarang serta peringkat indeks kesukaran.

Hal ini merupakan suatu keharusan yang jarang dilakukan, guru dalam penyelenggaraan hasil belajar terkadang mendapat kesulitan dalam penyelenggaraan tes formatif dan sumatif, permasalahannya adalah guru memiliki kelas paralel yang cukup banyak, sebagai contoh dalam pelaksanaan tes formatif yang diselenggarakan sesuai dengan jam pembelajaran masing-masing jika hanya menggunakan satu instrumen saja tidak menutup kemungkinan saling menyontek dan bisa terjadi kebocoran soal dengan kelas yang lainnya. Maka untuk menghindari situasi demikian guru biasanya hanya mengambil perangkat tes Ujian Nasional yang ada. Sedang soal Ujian Nasional yang digunakan tidak ada informasi sebelumnya bahwa paket-paket soal tersebut sudah disetarakan.

Dengan demikian penilaian guru terhadap peserta didik tidak mengukur yang sebenarnya. Misalnya peserta didik mendapat nilai 90 karena mengerjakan tes yang kategori mudah sedang peserta didik yang mendapat nilai 50 karena mengerjakan soal yang kategori tingkat kesulitannya tinggi. Dari hal tersebut terjadi penilaian yang tidak adil antara peserta didik yang satu dengan yang lainnya. Sehingga sangat diharapkan agar melakukan penyetaraan terutama bagi pendidik sebelum memberikan tes yang berbeda terhadap peserta didik dengan paket soal yang berbeda sehingga adil dalam memberikan penilaian. Penyetaraan tes perlu dilakukan khususnya bagi kegiatan pengujian dalam skala besar dalam mempersiapkan lebih dari satu perangkat tes mengingat bahwa menyusun tes yang benar-benar paralel tidaklah mudah. Dengan demikian, menjadi suatu keharusan bagi pengembang tes atau lembaga pengembang tes untuk melakukan penyetaraan terhadap perangkat tes yang digunakan.

Berdasarkan permasalahan tersebut maka peneliti akan mengadakan penelitian tentang penyetaraan tes. Dengan harapan soal yang setara bisa dijadikan sebagai bahan tes, misalnya dalam pelaksanaan ulangan harian, sehingga tingkat kesadaran untuk menyontek tidak terjadi lagi. Karena peserta didik tidak merasa ada diskriminan dalam menyelesaikan tes yang berbeda. Hal yang berkaitan dengan penyetaraan seperti panjang tes, karakteristik dari sekumpulan anchor item, ukuransampel minimal, variasi kemampuan peserta tes, metode penyetaraan, kesalahan penyetaraan sangat penting untuk diteliti dalam kerangka teoritis dan keperluan pengukuran di Indonesia.

Secara teori berdasarkan hasil penelitian semakin besar ukuran sampel, semakin stabil hasil estimasi parameter butir dan kemampuan. Hasil estimasi parameter butir dan kemampuan terlibat langsung dalam perhitungan konstanta penyetaraan. Akibatnya kestabilan metode penyetaraan dipengaruhi oleh ukuran sampel. Semakin besar ukuran sampel, semakin stabil metode penyetaraan yang digunakan. (Kartono, 2006: 60)

Menurut Syaifuddin (2005) dalam penelitiannya menggunakan ukuran sampel 300, 500, dan 800 menyatakan bahwa “ukuran sampel memiliki pengaruh langsung pada keakuratan estimasi parameter item tes, untuk itu sampel representatif dan besar bisa diharapkan memberikan hasil penyetaraan yang lebih akurat”. Pada penelitian ini mengkaji tentang pengaruh ukuran sampel terhadap hasil penyetaraan tes ujian nasioanl matematika SMP berdasarkan teori respon butir. Berbagai penelitian tentang ukuran sampel, yaitu ukuran sampel memiliki pengaruh langsung pada keakuratan estimasi parameter item tes, demikian juga pada tingkat kesalahan penyetaraan. Kim & Cohen (2000) melakukan penelitian tentang perbandingan *linking* dan *concurrent calibration* model respon berjenjang dengan ukuran sampel yang digunakan 300 dan 1000, hasil penelitian menunjukkan keakuratan penyetaraan tergantung dari ukuran sampel. Selanjutnya Reise & Yu(1990) dalam Syaifuddin (2005) melakukan penelitian tentang perbaikan parameter estimasi pada model respon berjenjang, ukuran sampel yang digunakan adalah 250, 500, 1000 dan 2000. Hasil penelitian menunjukkan semakin besar ukuran sampel yang digunakan, korelasi antara *true parameter* dengan parameter estimasi semakin tinggi.

Swediati (1997) melakukan penelitian tentang penyetaraan tes pada *Generalized Partial Credit Model (GPCM),* ukuran sampel yang digunakan adalah 500 dan 1000. Hasil penelitian menunjukkan bahwa keakuratan penyetaraan meningkat seiring dengan meningkatnya ukuran sampel. Penelitin yang dilakukan oleh Bastari. (2000) dalam Syaifuddin (2005) tentang *lingking multiple-choice and constructed-response items to a common proficiency scale.* Penelitian dilakukan untuk menyelidiki pengaruh panjang tes, proporsi *item* pada setiap naska tes, panjang *anchor test,* ukuran sampel, perbedaan distribusi kemampuan kelompok, dan metode penyetaraan, ukuran sampel yang digunakan untuk pembangkitan data 400, 1000, dan 3000. Hasil penelitian menunjukkan ukuran sampel mempunyai pengaruh terhadap kualitas penyetaraan. Semakin besar ukuran sampel yang digunakan pada penyetaraan semakin akurat hasil penyetaraan. Beragamnya ukuran sampel yang digunakan bertujuan untuk menemukan ukuran sampel minimum . Umumya penyetaraan dilakukan untuk suatu sistem penilaian berskala besar. Oleh karena itu hasil penelitian berkenaan dengan ukuran sampel kecil sangat membantu praktisi penilaian karena penyetaraan menjadi lebih efisien, menghemat biaya, dan terapan/ penyetaraan di sekolah dapat dilakukan. Eid dalam Rijanto (2012) meneliti pengaruh sampel (*sample size*) pada penyetaraan butir tes mengusulkan untuk menggunakan ukuran sampel 200,400,dan 800. Setiadi. (1997) dalam Tulimisar (2006) dalam penelitiannya terhadap estimasi parameter butir menyatakan sampel yang relatif kecil berukuran 100 atau 200. Sedangkan Livingston & Feryok (1987) dalam Rijanto (2012) melakukan penelitian pada penyetaraan ekipersentil estimasi frekuensi dengan penghalusan pada sampel berukuran 100 sampai dengan 3000, dan akurasi penyetaraan terjadi pada sampel berukuran 300. Kolen & Brennan (2004) dalam Kurtz dkk (2013) tentang penyetaraan sampel kecil dengan menggunakan SAS Macro dengan ukuran sampel kecil 400. Tulimisar (2006) mengadakan penelitian tentang Akurasi relatif penyetaraan skor tes ditinjau dari metode penyetaraan dan teknik penghalusan menyimpulkan bahwa akurasi relatif penyetaraan skor tes untuk sampel berukuran 300. Berdasarkan beberapa penelitian tersebut maka penelitian ini menggunakan ukuran sampel 200, 400, dan 700 sebagai perbandingan apakah ukuran sampel 200, 400, dan 700 ada perbedaan keakuratan penyetaraan dari dua perangkat tes yang berbeda. .

1. Rumusan Masalah

Berdasarkan latar belakang permasalahan di atas, maka rumusan masalah dalam penelitain ini adalah :

1. Apakah terdapat perbedaan hasil penyetaraan tes ujian nasional matematika SMP berdasarakan ukuran sampel 200 dan 400?
2. Apakah terdapat perbedaan hasil penyetaraan tes ujian nasional matematika SMP berdasarkan ukuran sampel 200 dan 700?
3. Apakah terdapat perbedaan hasil penyetaraan tes ujian nasional matematika SMP berdasarkan ukuran sampel 400 dan 700?
4. Tujuan Penelitian

Berdasarkan rumusan masalah maka tujuan penelitian ini adalah untuk mengetahui :

1. Perbedaan hasil penyetaraan tes berdasarkan ukuran sampel 200 dan 400.
2. Perbedaan hasil penyetaraan tes berdasarkan ukuran sampel 200 dan 700.
3. Perbedaan hasil penyetaraan tes berdasrakan ukuran sampel 400 dan 700.
4. Manfaat Penelitian

Manfaat yang diharapkan dari pelaksanaan penelitian ini adalah sebagai berikut :

1. Teoritis
2. Bagi Peneliti yang lain, terutama menyangkut penyetaraan perangkat tes, akan menambah khasanah empirik dalam rangka pengembang teori yang berkaitan dengan evaluasi dan tes hasil belajar.
3. Bagi peneliti pribadi, hasil penelitian ini dapat digunakan sebagai bahan penelitian lanjutan, khususnya pada model dan metode yang lain, berpengaruh pada proses penyetaraan tes.
4. Praktis
5. Bagi guru, dapat digunakan sebagai bahan masukan, mengenai perlunya penyetaraan skor dari dua tes atau lebih yang berbeda, sehingga adil dalam menilai hasil belajar peserta didik.
6. Bagi peserta didik, akan merasa memperoleh perlakuan yang lebih adil, karena mendapatkan tes yang setara, sehingga tidak merasa dirugikan atau diuntungkan.
7. Bagi sekolah, sebagai bahan masukan untuk dapat mengembangkan bank soal dan pelaksanaan tes dalam evaluasi hasil belajar, terutama menyangkut penyetaraan tes.